Characterising functional diversity in protein domain superfamilies and metagenomes

Natalie Louise Dawson

A thesis submitted for the degree of

Doctor of Philosophy

February 2015

Institute of Structural and Molecular Biology

University College London

I, Natalie Louise Dawson confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Natalie Louise Dawson

February 2015

Abstract

The majority of CATH domain structure superfamilies have small populations and are conserved in sequence and function. However, previous studies have shown that $\sim 4\%$ are highly populated and functionally diverse. Previous analyses of some of these showed that relatives with different functions tend to exploit different functional sites to perform their function. In this work, functional site diversity was explored with a much larger dataset of superfamilies, by examining residues involved in protein interfaces and catalytic sites. This was done using a novel protocol to map sites across each superfamily. Functional site locations were shown to be least diverse for catalytic sites and most diverse for protein-protein binding sites. However, although protein interaction sites can vary considerably, in 79% of superfamilies analysed there is a common protein interface site, used by at least 80% of the functionally diverse relatives.

By contrast with protein interactions, enzyme superfamilies tend to use the same active site in functionally diverse relatives. However, sometimes the nature and location of catalytic residues vary. We examined changes in catalytic machinery over one hundred enzyme superfamilies by considering physicochemical properties and sequence/structure positions. Reaction mechanisms were also compared to explore how enzyme chemistry has evolved between functionally diverse relatives and how changes in chemistry relate to changes in catalytic residues. A complex relationship was found and several examples are discussed to illustrate the different trends identified.

In the final chapter, we assigned metagenome sequences to functional families in CATH and used KEGG pathway annotations to identify differences in the functional abilities of two metagenome environments, the human tongue and gut. Bacteroidetes, Firmicutes, and Proteobacteria phyla dominate both microbiomes. Enriched functional terms in the tongue and gut environments suggested an enrichment of bacterial cell wall building proteins in the mouth and an enrichment of denitrifying enzymes in the gut.

Acknowledgements

I would like to thank the Institute of Structural and Molecular Biology at UCL/Birkbeck for accepting me onto the Wellcome Trust PhD Interdisciplinary PhD Programme and the Medical Research Council for my studentship.

A very big thank you to my PhD supervisor, Professor Christine Orengo, for her guidance, patience, and enthusiasm. I would also like to acknowledge my secondary supervisor, Professor John Ward, and my thesis committee chair, Dr. Andrew Martin for their support.

Thanks are also in order to past and present members of the Orengo group. In particular, thanks to Ian Sillitoe, Tony Lewis, Benoit Dessailly, Romain Studer, Corin Yeats, Jon Lees, and Jim Perkins for their help and advice with any computational or programming questions/issues. Thanks also to Benoit Dessailly for helpful discussions about functional diversity in protein domain superfamilies. Thanks to Robert Rentzsch, Dave Lee, and Sayoni Das for helpful discussions on the subclassification of CATH superfamilies into functional families. Thanks also to Dave Lee for helpful discussions on comparative metagenome analysis, and to Ali Cuff for useful discussions on the CATH classification process. Thanks to Anja Baresic and Nouf Alnumair, previous members of the Martin group, for their help and advice over the last four years. Finally, a big thank you to everyone in the office for the almost constant supply of cake, bread, and sweets!

There are a number of people outside of university that have been truly great in providing me with emotional support throughout my time at UCL. First of all I would like to thank my family, in particular my Mum, Dad, Grandma Croft, and sister, Emma. Also, thanks to my good friends Anathe Patschull, Katherine Tomlinson, Kirstyn Twumasi, Elizabeth Barratt, Liz Dawson, and Rebecca Wood. Last, but not certainly not least, the biggest thanks goes to Oliver Willhöft for believing in me, and always being there for me.

Contents

List of Figures

List of Tables

List of Abbreviations

BLAST Basic Local Alignment Search Tool

COMPASS COmparison of Multiple Protein Alignments with Assessment of Statistical Significance

CSA Catalytic Site Atlas

DeconSeq DECONtamination of SEQuence data

DNA DeoxyriboNucleic Acid

EBI European Bioinformatics Institute

EC Enzyme Commission

EMBL European Molecular Biology Laboratory

GO Gene Ontology

HMM Hidden Markov Model

IBIS Inferred Biological Interactions Server

KEGG Kyoto Encyclopedia of Genes and Genomes

MACiE Mechanism, Annotation, and Classification in Enzymes

MIRA Mimicking Intelligent Read Assembly

MSA Multiple Sequence Alignment

NCBI National Center for Biotechnology Information

NGS Next Generation Sequencing

OTU Operational Taxonomic Unit

PDB Protein Data Bank

PRINSEQ PReprocessing and INformation of SEQuences

PSSM Position-Specific Scoring Matrix

rRNA ribosomal RiboNucleic Acid

SCI-PHY Subfamily Classification In PHYlogenomics

SDP Specificity Determining Position

SFF Standard Flowgram Format

SFLD Structure Function Linkage Database

SSG Structurally-Similar Group

Chapter 1

Introduction

Proteins have evolved to perform thousands of different functions. To understand this process of evolution we can study their amino acid sequences and their structures (where available) to find clues on their evolutionary relationships. Furthermore, once we understand the link between a protein's structure and sequence, and the function it performs, we can use this knowledge to predict the protein function of new uncharacterised proteins which are related to the ones studied. With the exponential rise of sequence data through fields such as metagenomics, the prediction of function is a highly important and valuable tool.

In this thesis, work is presented which describes the evolution of functions in many different protein families in the CATH classification of domain families. We then use in-house function annotation tools to predict the functional repertoires in the diverse microbial communities found in the human mouth and gut.

1.1 Expansion of protein sequence data

As previously mentioned, sequenced DNA provides a wealth of information needed to build an understanding of what an organism can do. The latest publication from the genomes online database (GOLD) (Pagani *et al.*, 2012) reported 6,908 complete genome sequencing projects and 12,972 incomplete projects. These are projects that are publicly available. With the evolution of sequencing technologies and the fall in the price of sequencing, the amount of sequence data is rising exponentially. The GOLD statistics for 2014 show that bacteria are the most sequenced kingdom of life with 36,413 genome projects, compared to 8,574 eukaryotic, 5,035 metagenomic, 4,381 viral, and 906 archaeal genome projects (see `https://gold.jgi-psf.org/statistics`).

Much of the protein sequence data is available from the Universal Protein Resource Knowledge Base (UniProtKB). UniProtKB/SwissProt (discussed in more de-

tail in Section 1.5) comprises 546,000 (in release 2014_07) manually-curated protein sequences and has doubled in size since $\sim$ 2008. UniProt/TrEMBL, representing the unreviewed section of UniProtKB, is growing even more rapidly and is currently at 79,824,243 sequences (in release 2014_07). TrEMBL has seen an extremely large increase in sequence data, doubling in size in less than every two years (The UniProt Consortium, 2014).

The National Centre for Biotechnology Information (NCBI) hosts two sequence databases, GenBank (Benson *et al.*, 2013) and RefSeq (Pruitt *et al.*, 2005). GenBank and RefSeq comprise protein and nucleotide sequence data, as compared to UniProt which contains only protein sequence data. The latest release of Genbank (version 203, August 2014) comprises 174,108,750 un-curated and redundant nucleotide sequences. Sequences from large-scale sequencing projects have also been added since April 2002. GenBank exchanges data daily with partners worldwide, the European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL-Bank) which is part of the European Nucleotide Archive (ENA) (Leinonen *et al.*, 2011) and the DNA Data Bank of Japan (DDBJ) (Kaminuma *et al.*, 2011), to provide a collection of comprehensive sequence data. RefSeq is a curated, non-redundant collection of nucleotide and protein sequences. Its entries are curated from redundant sequences in GenBank and it is currently limited to organisms with completely sequenced genomes.

1.2 Expansion of protein structure data

Solved three-dimensional protein structures obtained by Electron Microscopy, NMR, and X-ray crystallography methods are deposited in the Protein Data Bank (PDB) (Bernstein *et al.*, 1977). Bernstein *et al.* (1977) reports that the first structures were deposited in 1976. Subsequently, tens to hundreds of structures were deposited each year until the 1990s which saw the start of the NIH-funded Protein Structural Initiative (PSI) projects and a large increase in the amount of structural data published (see Figure 1.1). Thousands of structures were deposited from the

mid-1990s through to tens of thousands from 1999 onwards. As of 2014, the number of total protein structures in the PDB has passed the 100,000 mark.

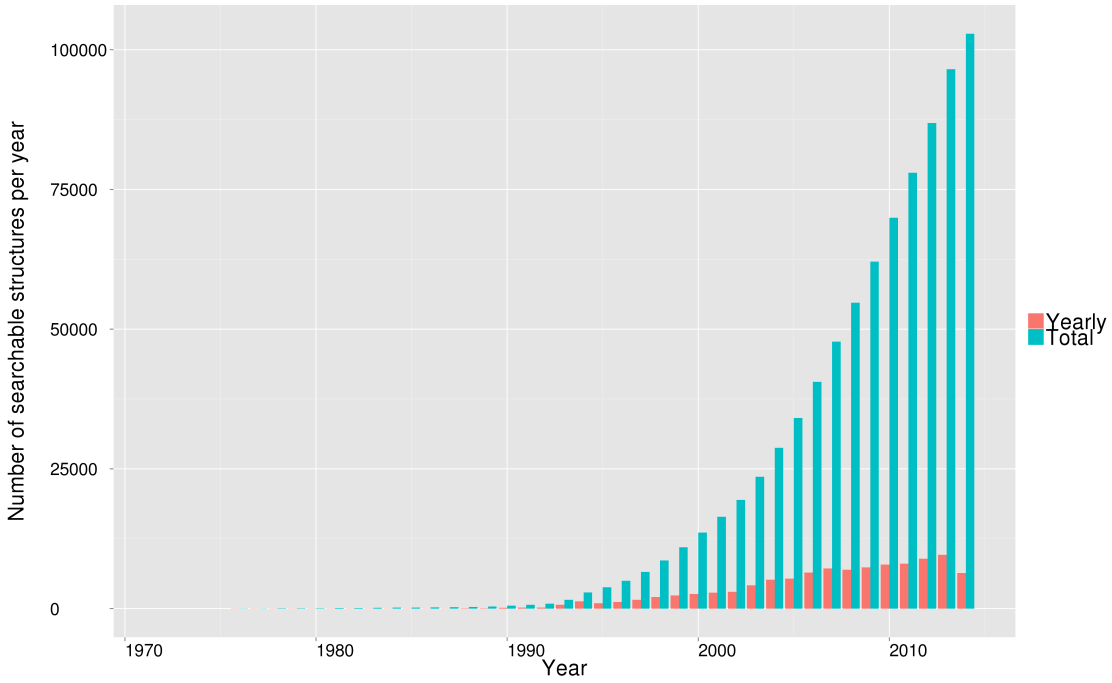


Figure 1.1: The yearly growth of the PDB from 1970 to 2014.

1.3 The organisation of sequence and structural data

Protein sequence and structure data has been organised into evolutionary families by a number of resources. Most notably, the CATH-Gene3D (Sillitoe *et al.*, 2013; Lees *et al.*, 2014) resource, and the SCOP (Andreeva *et al.*, 2014) and SUPERFAMILY (Wilson *et al.*, 2009) resources. Protein structures from the PDB are classified into evolutionary related groups, or superfamilies, in the CATH and SCOP resources, and protein sequence data are added to these superfamilies through the Gene3D and SUPERFAMILY resources.

Proteins that share the same common ancestor are known as homologues. This differs from anatomical homology which can describe the evolutionary relationship between two or more structures e.g. limbs, or even species. Homologous proteins can

be further described as orthologous or paralogous. Two proteins are orthologues, or are orthologous, if they are found in different species and have descended from a single gene in their last common ancestor. Paralogues on the other hand occur within a genome and are genes resulting from the duplication of a single gene (Sonnhammer and Koonin, 2002).

Protein domains are independently-folding functional and evolutionary units of a protein sequence that form the building blocks of the protein structure. The CATH database (established in 1997, Orengo *et al.* (1997)) is a hierarchical classification of protein domain structures. To classify protein domains in CATH, protein structures from the PDB are first split into separate chains. Domain boundaries are automatically inferred if there is enough similarity to existing domains, based on a sequence or structure similarity to already classified domains in CATH. The structural comparison program, CATHEDRAL (Greene *et al.*, 2007) is used to detect structure similarity. If there is not enough similarity, domain boundaries are manually defined, i.e. the protein chain sequence is cut at the identified domain boundaries. To detect sequence similarity, sequences of new PDB chains are scanned against a library of hidden Markov models (HMMs) from each CATH superfamily. The scores obtained from the CATHEDRAL and HMM scans are used to determine whether the domains identified in a new PDB chain are homologous to any domains classified in CATH. If there is enough evidence they are automatically classified into a superfamily. If not they are manually classified, for example, domains that are remote homologues typically have to be manually classified (Greene *et al.*, 2007).

At the top level of the CATH hierarchy, the Class (or C-level) is used to classify domains based upon their secondary structure content (see Figure 1.2). Class 1 domain structures are mostly alpha-helical, Class 2 domain structures are mostly beta-sheet, Class 3 domain structures have significant amounts of alpha-helical and beta-sheet structural elements, and Class 4 domains have very little secondary structure (Sillitoe *et al.*, 2013).

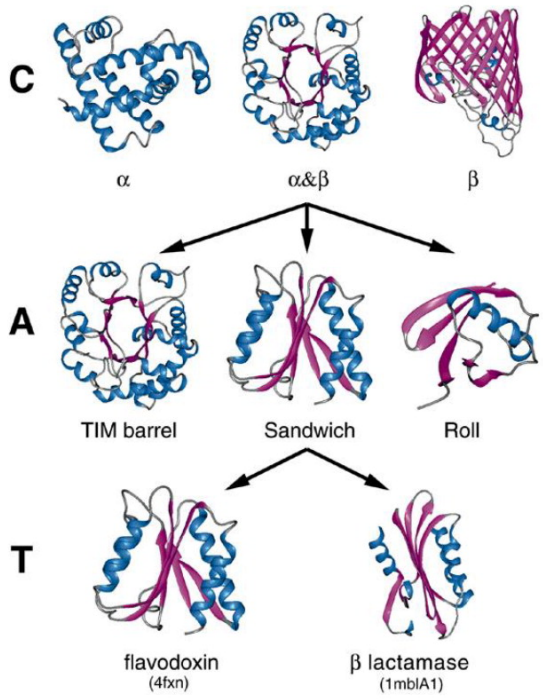


Figure 1.2: The first three levels of the CATH classification. Taken from Orengo *et al.* (1997).

The second level of the hierarchy is the Architecture (A-level), which represents similarities in how the secondary structures are arranged in 3D space. The third level is the topology or fold groups (T-level), which captures both the 3D arrangement and the connections between the secondary structures. Finally, the fourth level, the homologous superfamilies (H-level) contains domain structures that have $> 35\%$ sequence identity, as well as sufficient structural and functional similarity. The latest version of CATH, v4.0, consists of 277,687 structural domains from 69,058 PDBs classified into 2738 homologous superfamilies (Sillitoe *et al.*, 2013).

The Structural Classification Of Proteins (SCOP) database (established in 1995), like CATH, classifies protein domains using structural and evolutionary relationship information. The top level, Class, also classifies by secondary structure content. The second level groups domains by their structural folds, which is more similar to the T-level in CATH. The third level of SCOP represents the classification into homologous superfamilies (comparable to the H-level in CATH). Although SCOP is largely based on manual curation, unlike CATH, the curators do use some automatic

methods e.g. DALI, Pfam, BLAST to check for homology (Andreeva *et al.*, 2014).

As the PDB has grown and more remote evolutionary relationships detected, the relationships between domain structures, in some fold groups and homologous superfamilies, have been found to be more complex than previously thought. A SCOP2 prototype has therefore been developed which still organises protein domains according to their structural and evolutionary relationships, but instead forms a network from these relationships rather than a simple hierarchy (Andreeva *et al.*, 2014).

The Gene3D resource (Lees *et al.*, 2014) adds protein sequences to CATH from UniProt and ENSEMBL for proteins with no structural data. These sequences are predicted to belong to CATH domain superfamilies by scanning them against the HMM libraries built for each superfamily. CATH is expanded many hundreds-fold by this approach and some of the sequences have experimental characterisation, which increases the functional information for each CATH superfamily. The latest version of Gene3D (version 12.0) consists of 25,615,754 CATH version 4.0 protein domain assignments from $>$ 6000 cellular genes from ENSEMBL and $>$ 20 million unique protein sequences from UniProtKB. This was a 45% increase in sequence data compared to the previous release. Pfam and SUPERFAMILY domain annotations have also been added to the resource in the latest release to expand domain sequence coverage (Lees *et al.*, 2014).

The SUPERFAMILY resource (Gough *et al.*, 2001; Wilson *et al.*, 2009) is similar to Gene3D as it uses structural domain classifications to identify domains within UniProt sequence data, however these classifications are from SCOP rather than CATH. HMMs based upon the SCOP domain classifications are used to identify domains and structurally annotate the sequence data. The latest publication of SUPERFAMILY reported protein domain assignments for $>$ 900 genomes (Wilson *et al.*, 2009).

By contrast, the Pfam resource (Finn *et al.*, 2014) is a domain family resource in which evolutionary relationships are captured mainly using sequence data. In Pfam,

each protein family is represented by two multiple sequence alignments (MSAs) and a profile HMM for the identification of Pfam domains in new sequences. One of the MSAs is the seed alignment, i.e. all the protein sequences used to define the family, which contains a small number of family representative members. The other MSA contains the full sequence alignment, comprising sequences for all family members. The sequences for the full alignment are taken from UniProtKB. Curated families are known as Pfam-A families. There are also automatically generated families to improve sequence coverage, known as Pfam-B families. As mentioned, Pfam families have largely been generated using sequence data, however new families added to the latest release (version 27.0) have been identified by also using structural data in the PDB and CATH. For example, where a CATH domain does not match a Pfam family, the CATH domain sequence is used to iteratively search Pfam sequence database to extract sequences for a new Pfam family seed alignment (Finn *et al.*, 2014).

Gene3D, SUPERFAMILY, and Pfam are part of the InterPro consortium, which comprises 11 databases (Hunter *et al.*, 2012). In addition to those described already, InterPro also contains the protein family resources PANTHER (Mi *et al.*, 2010), PIRSF (Nikolskayaw *et al.*, 2006), TIGRFAMs (Selengut *et al.*, 2007), and HAMAP (Lima *et al.*, 2009), and the protein domain family resources PROSITE (Sigrist *et al.*, 2010), PRINTS (Attwood *et al.*, 2003), ProDom (Bru *et al.*, 2005), and SMART (Letunic *et al.*, 2009).

These databases provide diverse information about protein families, domains, functional site data, and conservation data. Each database is manually integrated into InterPro so that each non-redundant protein family has a combination of models based upon the different sources. The accuracy of each family is also manually checked by exploring the functional annotations such as GO terms, family HMMs or other types of representative model produced, and family names. One can scan a sequence of interest against all members of InterPro using the online tool or standalone program, InterProScan (Jones *et al.*, 2014) to obtain domain and functional

annotations.

1.4 The definition of protein function

It is difficult to define function as it is a term that only makes sense in context. Furthermore, although a lot of research has been conducted into the function of proteins, the functions of other entities, e.g. non-coding RNA and organelles (Bork *et al.*, 1998) are generally less well characterised. However, we shall focus on proteins as these play a large role in defining the functions of phenotypes for a particular organism. There are numerous areas with which function can be associated, ranging from biochemical function (e.g. catalytic activity) to biological processes and pathways (e.g. metabolic pathways, signal cascades), through to organ and organism function (e.g. physiology, behaviour). When annotating protein function it is therefore possible to describe function at any one of these levels. This can make it hard to compare function between different proteins as they may be annotated at different levels. The terms used can differ between researchers and unless the protein's functional annotation, or a close homologue, has been experimentally confirmed it may be difficult to trust.

For many years, most functional annotations were written as free text in the public literature and they contained a wide range of terminology and synonyms. Recent years have seen the improvement of natural language processing and the extraction of information from the literature, however a much bigger step has been the creation of standards in defining function (Lee *et al.*, 2007). One of the first standards was the Enzyme Commission (EC) numbering scheme, which is used to classify enzyme function (Barrett, 1992). There are many other important resources that provide standardised functional categories. For example, the Clusters of Orthologous Groups of proteins (COGs) (Tatusov *et al.*, 2003); ENZYME (Bairoch, 2000) which, uses EC numbers in its annotations; the manually-curated Swiss-Prot database (Wu *et al.*, 2006); the Functional Catalogue (FunCat) from the Munich Information Centre for Protein Sequences (Ruepp *et al.*, 2004); the Kyoto Encyclo-

pedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2014), a biological systems database that integrates molecular biology with systemic information; and MetaCyc (Caspi *et al.*, 2010), a database of over 900 experimentally-confirmed metabolic pathways. Controlled vocabularies have also been developed, for example the Gene Ontology (GO) (Ashburner *et al.*, 2000) schema (described more below), which is a relatively recent standard that is open source and provides the user with three types of structured controlled vocabulary with which to annotate protein sequences.

1.5 Protein function data resources

This section describes in more details the different resources used to provide protein function information. Those resources that focus only on enzyme data will be discussed in Section 1.5.1.

The GO is the most comprehensive resource providing information on protein functions. It provides a standardised set of functional terms for annotating protein sequence data which comprises three functional term ontologies: biological processes (BP), molecular function (MF), and cellular compartment (CC). The BP ontology consists of terms that refers to a protein's biological objective or the pathway or process that it is involved in. The MF ontology comprises terms related to the molecular function of the protein and is intended to capture conceptual protein function categories. One or more ordered assemblies of MF terms can be used to describe a biological process. The CC ontology consists of terms related to the cellular location of the protein when it is performing its function (Ashburner *et al.*, 2000). GO terms in these ontologies are connected as nodes in a directed acyclic graph where one term can have a relationship with one or more other terms. This allows the ontology networks to be easily updated as more information is gathered. It also means that the ontology networks can be flexible in reflecting the different ways in which different organisms perform a given function, for example the nucleus is re-arranged in different ways by different organisms during mitosis (Ashburner *et al.*, 2000).

GO categories have been linked to sequence data in numerous resources to provide standardised protein function annotations. These resources include: CATH, SCOP, Pfam, ENZYME, FunCat, and SwissProt (Ashburner *et al.*, 2000).

The SwissProt resource represents the reviewed section of UniProtKB which contains manually annotated protein entries. Information is extracted from the literature and computational analysis, and the curators aim to annotate protein family representatives across a diverse range of taxonomic groups. While there are many data extracted from the literature, those related to function include: the protein function, catalytic activity, pathway information, and functional site residue data (The UniProt Consortium, 2014).

The FunCat database (Ruepp *et al.*, 2004) is a hierarchically structured classification and also uses its own controlled vocabulary. It was first developed to describe the biology of yeast but has since been extended to the plant Arabidopsis, prokaryotes, and animals. Functional classification is based on seven main categories: metabolism, information pathways, transport, perception and response to stimuli, developmental processes, localisation, and experimentally uncharacterised proteins (Ruepp *et al.*, 2004).

1.5.1 Enzyme function data resources

There are thousands of different enzyme functions, which are typically classified using their chemical reaction information through the hierarchical Enzyme Commission (EC) numbering system (Barrett, 1992).

EC numbers are manually assigned to enzymes using published experimental data by the Joint Commission on Biochemical Nomenclature (JCBN) of the International Union of Biochemistry and Molecular Biology (IUBMB) and the International Union of Pure and Applied Chemistry (IUPAC) (Kotera *et al.*, 2004). There are four (numbered) hierarchical levels to an EC number: the first level represents the class of the catalysed reaction, which comprises six categories (see Table 1.1); the second level (the sub-class) represents the bonds the reaction acts on; the third level

(the subsub-class) represents the chemical functional group acted on; the fourth and final level represents the substrate for the enzyme-catalysed reaction. There are currently 5445 active EC entries (09-Jul-14 release) in the ExPASy ENZYME database (Bairoch, 2000).

EC Class Number	Name	Description
EC 1	Oxidoreductases	Enzymes that catalyse the transfer of electrons from one molecule (the reductant, or the electron donor) to another molecule (the oxidant, or the electron acceptor)
EC 2	Transferases	Enzymes that transfer a functional group from one compound (the donor) to another (the acceptor)
EC 3	Hydrolases	Enzymes that catalyse the hydrolysis of various bonds
EC 4	Lyases	Enzymes that cleave bonds in ways other than hydrolysis or oxidation. Two substrates are usually involved in one reaction direction, however only one in the other direction
EC 5	Isomerases	Enzymes that catalyse the conversion of one isomer to another
EC 6	Ligases	Enzymes that catalyse the joining of two molecules through the formation of a new chemical bond

Table 1.1: Description of each Enzyme Commission (EC) class (Barrett, 1992).

Enzyme-related data are also available from a number of online database resources. For example, the BRaunschweig ENzyme DAtabase (BRENDA) enzyme portal (Schomburg *et al.*, 2013) contains functional biochemical and molecular enzyme data and currently consists of 2.7 million semi-curated entries detailing enzyme occurrence, function, kinetics, and molecular properties (Schomburg *et al.*, 2013). BRENDA is a large resource that also includes enzyme-related information from automatic text mining and from analysing relationships between EC number and disease.

The KEGG ENZYME database (Kanehisa *et al.*, 2014) is based upon the ExplorEnz database (McDonald *et al.*, 2009), which uses the IUBMB EC nomenclature

to define enzyme function.

The integrated relational enzyme database, IntEnz (Fleischmann *et al.*, 2004) is the official version of the Enzyme Nomenclature and contains the most recent version of the EC number list, decided by the Nomenclature Committee of the IUBMB. The aim of this database is to create a single relational enzyme database linking three data sources: 1) the Enzyme (EC) List; 2) the Enzyme Nomenclature database, ENZYME; and 3) the enzyme function database, BRENDA.

1.5.1.1 Reaction mechanism data

The Mechanism, Annotation and Classification in Enzymes (MACiE) database (Holliday *et al.*, 2011) consists of entries describing enzyme reaction mechanisms and contains information on similarity of reaction pairs. Each reaction mechanism entry includes the chemical reaction diagram, a walk-through of each reaction step, and the catalytic residues involved. In the comparison of reaction mechanisms, the similarity of a reaction pair depends on the similarity of the reaction steps in each mechanism (O'Boyle *et al.*, 2007).

There is a tool available in MACiE for the comparison of reaction mechanism entries, which calculates the similarity of reaction mechanisms for enzyme A and B using a similarity score calculated using the Tanimoto coefficient (Equation 1.1). Reaction information is stored in bitmaps, where each bit represents a modelled characteristic. In an enzyme object a characteristic can either be present, i.e. 1, or not present, i.e. 0. The variable c represents the total number of common bits between A and B that have a value of one. The variables a and b represent the total number of bits that have a value of one in each A and B, respectively. This method is used to calculate similarity in catalytic machinery and reaction mechanism (O'Boyle *et al.*, 2007).

$$T(A, B) = \frac{c}{a + b - c} \tag{1.1}$$

SABIO-RK (Wittig *et al.*, 2012) is a manually-curated database that takes in-

formation from the literature, storing standardised biochemical information and corresponding kinetic properties. It is a quantitative reaction-orientated source of information, which includes rate equations, kinetic parameters, and the experimental and environmental conditions used to calculate the kinetics.

The KEGG REACTION resource contains all reactions from KEGG ENZYME and additional reactions from the metabolic pathways in KEGG PATHWAY (Kanehisa *et al.*, 2014) (see below for more details on this).

1.5.1.2 Metabolic pathway data

The Reactome resource is an open-source database that focusses on human biological metabolic pathways. Croft *et al.* (2014) reported functional information for 7088 human proteins (34% of the predicted human proteome) which take part in 6744 reactions and 1481 metabolic pathways.

The MetaCyc database (Caspi *et al.*, 2010) contains experimentally determined metabolic pathways that are curated from the literature, and enzyme data. There are more than 1800 pathways from more than 30,000 publications, which make it the largest collection of metabolic pathways available to date.

KEGG PATHWAY contains interactive maps of manually-drawn pathways that contain information on molecular interactions and chemical reactions. There are hundreds of pathways, which are grouped into seven main categories: metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases, and drug development. Each node in a pathway represents a set of orthologues (KEGG Orthologues, KOs) performing the same function. Organism-specific metabolic pathway maps have been generated as well as general overviews for all organisms (Kanehisa *et al.*, 2014).

1.5.1.3 Catalytic residue data

The Catalytic Site Atlas (CSA) (Porter *et al.*, 2004; Furnham *et al.*, 2014) is a resource providing catalytic residue annotation for enzymes in the PDB. There are two

types of data: a manually-curated set of entries taken from enzymes described in the scientific literature, and an automatically generated data set of catalytic residues mapped onto homologous structures of the manual entries. Homologues to the manual entries are identified in the PDB and SwissProt using SSEARCH36 (Sierk, 2004) using a strict statistical significant threshold, E-value, of $1x10^{-06}$. In the latest version of the CSA there are 968 manually-curated entries/structures, covering $\sim 30\%$ of the PDB, and 32,216 structures annotated using homology information (Furnham *et al.*, 2014).

1.6 The evolution of protein functions

1.6.1 Selective evolutionary pressures shaping the evolution of function

The duplication of genetic material provides the necessary material for sculpting additional and/or novel functions. For example, whole-genome duplication (polyploidisation) provides duplicate chromosomes, duplicate genes and duplicate regulatory regions driving gene expression (Force *et al.*, 1999). The first model of how products of gene duplication are used in protein function evolution was proposed by Ohno (1970). Based on the fact that gene duplication provided extra, redundant copies of genes, Ohno founded the classical function evolution model. The model hypothesises that, whilst one copy of the gene is under selective constraints to maintain the original function, the other copy of the gene is shielded by the first and is therefore free to evolve a novel function through amino acid mutations. This accumulation of mutations would either result in a non-functional pseudo-gene (non-functionalisation), due to the accumulation of deleterious mutations, or it would lead to the evolution of a novel function (neo-functionalisation), due to the accumulation of beneficial mutations (Ohno, 1970).

Lynch *et al.* (1998) found that deleterious mutations are more common than beneficial mutations and so the classical model predicts that it is more likely for

the gene copy to become a pseudo-gene. Force *et al.* (1999) later proposed a new model that proposed the function of duplicated genes would be preserved rather than lost. This model suggested that chromosomes and genes were preserved, i.e. kept functional, through subfunctionalisation. For example, if different genes on two duplicate chromosomes became functionally inactive (e.g. through gene inactivation, or loss of regulatory elements), for the organism to remain viable the two chromosomes must between them retain functional copies of all the genes found in the ancestral chromosome. In two duplicated genes where different regulatory elements are inactivated in each copy, it was proposed that the two genes had to be expressed at the same time to maintain the ancestral function and remain viable. The model therefore proposes subfunctionalisation as the process whereby duplicate genetic copies have each lost some functional aspect of the ancestral locus, however the function provided by the ancestral locus is still required and so both copies must work together to maintain this function (Force *et al.*, 1999).

While these phenomena have been frequently studied at the whole protein level in the past, such studies can mask the effects of mutations at the domain-level (Khaladkar and Hannenhalli, 2012). For example, it may be that a single, small domain within a multi-domain protein is evolving at a faster rate (asymmetric evolution) to the other domains, however the signal is lost within the signal-to-noise ratio. In cases of subfunctionalisation where two copies of the gene have domains that are evolving at different rates, these signals of varying evolutionary rates could cancel each other out at the whole-protein level. Therefore, studying the evolutionary rates of each domain allows a more accurate analysis and biological interpretation (Khaladkar and Hannenhalli, 2012).

Khaladkar and Hannenhalli (2012) demonstrated this with their large-scale study on duplicate genes in five teleost fishes. They found that previously observed asymmetry of overall protein evolution is mainly a result of the divergence of specific protein domains, reflecting the importance of studying evolutionary mutations at the domain level. Positive selection has also been found to contribute to the diver-

gence of function within duplicated genes. This is where the rate of non-synonymous mutations is greater than the rate of synonymous mutations. A number of studies (Zhang *et al.*, 1998; Duda and Palumbi, 1999; Hughes *et al.*, 2000) showed that non-synonymous amino acid mutations, i.e. mutations at the DNA level that lead to a change in amino acid, are favoured and therefore get fixed at a faster evolutionary rate than synonymous mutations, i.e. mutations at the DNA level that do not cause a change in amino acids.

Ohno's model assumption that the newly duplicated copy of the gene is not under selective pressure has been a subject of much discussion as duplication is not observed to be a frequent event (Khersonsky and Tawfik, 2010). Due to the extra, duplicated, copies of messenger RNA (mRNA) and protein being transcribed and translated, there is extra stress on the cell to produce more energy, which has been shown to induce selective pressure to inactivate these extra copies (Cooper and Lenski, 2000; Dekel and Alon, 2005; Wagner, 2005; Stoebel *et al.*, 2008). Gene duplication is however, frequently observed to be a positively-selected event when there are demands for higher amounts of a given protein (McLoughlin and Ollis, 2004; Bergthorsson *et al.*, 2007).

Studies have found that over one-third of the random mutations in a protein are deleterious (Bershtein *et al.*, 2006; Camps *et al.*, 2007; Tokuriki *et al.*, 2007) and only $\sim 10^{-03}$ of random mutations are thought to be beneficial (Khersonsky and Tawfik, 2010). In the cases where the duplicated copy is drifting without any selective pressures, it is orders of magnitude more likely to lose all functionality due to mutations affecting the protein folding and stability (non-functionalisation) rather than acquire new functionality (neo-functionalisation) (Bershtein and Tawfik, 2008). This is due to mutations typically affecting the folding and the stability of the protein (Yue *et al.*, 2005; Tokuriki and Tawfik, 2009).

Khersonsky and Tawfik (2010) therefore offer a number of revisions to Ohno's model, including: 1) a gene sharing model, where one gene is recruited for a different, moonlighting function without any changes in the DNA; 2) new protein functions

evolve through promiscuous intermediates without the need for prior gene duplication, gene duplication then allows the different functions of the intermediates to be optimised (Glasner *et al.*, 2006); 3) gene duplication undergoes positive selective pressures, increasing functional variability as the secondary functions of promiscuous proteins are selected for when present in high doses (Khersonsky and Tawfik, 2010).

1.7 The prediction of protein function

With the ever-increasing number of new sequences being determined from experimental efforts such as high-throughput next-generation sequence methods, it is physically impossible to perform all of the experiments needed to determine the functionality of each sequence. It is estimated that only $\sim$ 1%, 7%, 10%, and 20% of proteins have been experimentally characterised in the model organisms *Caenorhabditis elegans*, *Mus musculus*, *Drosophila melanogaster*, and *Homo sapiens* (Lee *et al.*, 2007) and computational methods are constantly being developed and improved to automatically predict enzyme function. In general, it is hard to predict protein function due to five main reasons: 1) as mentioned already, function can be defined from different perspectives and at a range of levels, for example the biochemical events a protein is involved in, or the role of an enzyme in a pathway; 2) a protein's function and its experimental characterisation are context-dependent, therefore there may not be experimental information for all possible conditions; 3) proteins are often multifunctional and promiscuous; 4) experimental functional annotations are error-prone due to experimental misinterpretation and curation issues; and 5) protein function is currently mapped to gene names, however this is confusing for potential isoforms of those genes that have different functionality (Radivojac *et al.*, 2013).

A protein's function is encoded by its amino acid sequence and is dependent on other factors such as post-translational modifications. Computational methods have therefore been developed to predict protein function and thereby guide experimental validation. This can greatly reduce the cost of experimental validation. Over the

past couple of decades numerous methods have been written to predict protein function using protein sequence and protein structure information.

1.7.1 Sequence-based methods for protein function prediction

The most common methods used for predicting function from sequence exploit a sequence's accession code or text-based searches, related to the gene name for example, or sequence-sequence comparison algorithms to match known sequences stored in a database. Resources for such searches are provided by the National Center for Biotechnology Information (NCBI) in the USA and by the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI). If searching for a protein accession code or a gene name, for example, does not obtain a result, the query sequence can be submitted to the Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990). This tool matches a DNA or protein sequence to a homologous sequence, where possible, in a chosen biological database (Altschul *et al.*, 1990). Many of the tools available from the NCBI or EMBL-EBI have been integrated to allow the user to combine multiple tools in the same search. InterPro (Hunter *et al.*, 2009) from the EMBL-EBI is one such example (discussed in more detail in Section 1.3).

A threshold is applied to all of these sequence-matching methods for confidence purposes. For example, various studies have shown that a suitable threshold is 40% pairwise sequence identity when inheriting the first three digits of an EC number. More than 60% sequence identity is required to inherit all four digits (Rost, 2002; Addou *et al.*, 2009).

The GOtcha method (Martin *et al.*, 2004) uses BLAST and the Gene Ontology (GO) to predict the function of an unknown gene sequence. GOtcha first compares a query sequence against seven well-annotated genomes using BLAST. The GO terms associated with the gene sequences most similar to the query are retrieved. The method then calculates the probability of each GO term (and their ancestral

terms) being associated with the unknown sequence. A single probability value is provided for each GO term, allowing the user to easily assess whether that functional term could be valid. This approach is valuable as methods such as BLAST provide multiple outputs that must all be assessed before a match can be assigned.

The Protein Function Prediction (PFP) method (Hawkins *et al.*, 2009) also uses a sequence-based approach to predict GO functional terms but the method is based upon the results of Position Specific Iterative (PSI)-BLAST searches rather than BLAST. PSI-BLAST (Altschul *et al.*, 1997) is a profile-based method that constructs a profile from an initial BLASTp search and then uses that profile to iteratively search for increasingly distant homologues. For each sequence that is hit by PSI-BLAST, its associated GO terms are each scored according to their expectation value (E-value) (see Equation 1.2). The E-value (E) represents the number of different sequence alignments, with a score of at least S, that are expected to be found by chance.

$$E = Kmn\ e^{-\lambda S} \tag{1.2}$$

Where m and n represent the lengths of two random sequences, S represents the maximal segment pair score (i.e. the segment of the two sequences that provides the highest similarity score) $\geqslant S$, and K and λ are two parameters used in the evaluation of the E-value. The GO terms associated with each PSI-BLAST sequence hit are each scored according to 1) how often a particular GO term is associated with retrieved hits that are similar in sequence and 2) the sequence identity of the sequences in 1) with the query (Hawkins *et al.*, 2009).

One key feature of PFP is that it retrieves functional information from highly divergent sequences hits returned by PSI-BLAST, i.e. those hits with high E-values (up to 125) that are not considered statistically significant. Such high E-values however mean that there will be high levels of error, i.e. false positives. Hawkins *et al.* (2009) showed that using sequence hits with an E-value $\geqslant 1$, their benchmark sequence coverage using PFP was double that of the top PSI-BLAST annotations.

Pairwise methods like BLAST are still very popular as they are the simplest type of method. However more sensitive profile/pattern-based methods, for example PROSITE (Sigrist *et al.*, 2002) and PRINTS (Attwood, 2002) are also used to detect homologues with similar functions.

A more powerful type of profile is the hidden Markov model (HMM), which is better able to handle residue insertions/deletions (indels) between homologues. While sequence profiles only report the amino acid frequency at each alignment position, HMMs also report the probabilities for insertions and deletions. HMMs also have a higher sensitivity than sequence profiles as they heavily penalise sequence hits that are matched by chance, in comparison to true positives that frequently have insertions and deletions at the same alignment positions (Söding, 2005).

These are usually constructed from an MSA of homologues sharing similar functions. For example, the widely used Pfam database provides a library of HMM models for each family. However not all families are functionally coherent (Lee *et al.*, 2007). InterPro, which has been described previously in Section 1.3, combines function predictions from PROSITE, PRINTS, Pfam, and eight other resources (Hunter *et al.*, 2012).

However, it is not always possible to predict function from a protein sequence because a match in a database may not exist, or the sequence may exist but it may have no function assigned and simply be described as a 'hypothetical protein'. Another disadvantage of this sequence-based approach is the difficulty in distinguishing between residues that are structurally conserved and those that are functionally conserved (Sadowski and Jones, 2009). For example, the patterns of conserved residues captured by the profile may represent positions important for structural stability or packing, rather than function.

1.7.1.1 Subclassification of protein domain superfamilies

Homologous superfamilies can sometimes be large and functionally diverse. For example, less than 100 ($< 4\%$) superfamilies in CATH are very diverse in sequence,

structure, and function. These superfamilies also account for a disproportionate number of domain sequences as they contain more than 40% of domains in CATH (Dessailly *et al.*, 2009; Cuff *et al.*, 2011). These superfamilies comprise two or more different EC terms, and up to > 100 different EC terms in some superfamilies (Sillitoe *et al.*, 2013). To understand the divergence of function across a superfamily, a superfamily can be clustered into subfamilies, whereby each subfamily contains sequences that code for a similar function. As each member of a subfamily has the same functional properties, the key functional residues should be highly conserved throughout the subfamily members. By comparing different subfamilies it is possible to observe how these conserved, key functional residue change in order to provide different functionalities. For example, the GeMMA protocol (Lee *et al.*, 2010) clusters sequences in a superfamily into families of functionally coherent relatives, or FunFams, using a hierarchical agglomerative clustering method. This method is discussed further in Section 2.2.1, page 59.

Abhiman and Sonnhammer (2005a) carried out a large-scale prediction of functional shift in enzyme protein families and benchmarked their methods against experimentally-annotated data. This was the first study benchmarked on such a large scale. Enzyme families taken from the Pfam database (Sonnhammer *et al.*, 1997) were divided into subfamilies using Bayesian Evolutionary Tree Estimation (BETE) (Sjölander, 1998) and a change in function was identified using two types of site: Conservation-Shift Sites (CSS), where a sequence position in an MSA is completely conserved, however the two subfamilies each have a different amino acid type; and Rate-Shifting Sites (RSS), where the amino acid residues in the different subfamilies are evolving at different rates. The definition of a change in function in this analysis was a change in EC number (Abhiman and Sonnhammer, 2005a).

Subfamily Classification In PHYlogenomics uses an algorithm (SCI-PHY) (Brown *et al.*, 2007) to automatically detect and create subfamilies using functional and evolutionary information. While GeMMA defines a tree of sequence-similarity based clusters which is then cut into subfamilies, SCI-PHY builds a hierarchical tree using

sequence information as well as Dirichlet mixture densities and relative entropy measures. Dirichlet mixture densities are used here to construct subfamily profiles. The relative entropy measure is used to calculate distances between subfamilies, which are used to build the tree. HMMs are then derived from the subfamily sequences.

Capra and Singh (2008) developed the method, GroupSim to subclassify a set of homologous proteins using specificity determining positions (SDPs). SDPs are alignment positions that determine a subfamily's functional specificity and they are therefore used to aid protein function prediction. SDPs allow one to understand the mechanism of functional diversity within a superfamily and also enable the assignment of specific functions to uncharacterised proteins. To predict SDPs between subfamilies/specificity groups, GroupSim first compares all pairs of aligned residues within and between the subfamilies, where all subfamilies are aligned in one MSA. The average physicochemical similarity of all residue pairs in a group is calculated for each subfamily in the alignment. The method rewards alignment positions where a single amino acid residue is conserved in one subfamily but where a different amino acid is conserved in the other subfamilies, or where it is not conserved in another subfamily. Such positions are predicted to be SDPs (Capra and Singh, 2008).

1.7.2 Structure-based methods

If there is no detectable sequence similarity to a protein with known function, structure-based methods can be used to find a functional analogue, e.g. using active site and/or catalytic residue information. If this is not possible, one can look for structural features such as pockets on the surface of the protein that are likely to be active sites. Methods using these techniques are discussed below.

Firstly, as with pairwise sequence comparison, the simplest approach to predict function from structure is to perform a pairwise structural comparison.

1.7.2.1 Protein structure comparisons

If two proteins show very high structural similarity along their entire amino-acid sequence, it is likely that they have the same or a very similar function (Lee *et al.*, 2007).

There are several publicly-available methods which perform fast pairwise comparisons of secondary structure elements. For example, GRATH (Harrison *et al.*, 2003), which uses graph theory to match a query structure to a structure in the PDB, and the Secondary Structure Matching (SSM) method (Krissinel and Henrick, 2004) provided by the PDB. However to obtain an accurate structural alignment, all of the residues need to be considered. This also means that the methods need to cope with insertions and deletions (i.e. indels) and various strategies have been developed for doing this. For example, the STAMP (Structural Alignment of Multiple Proteins) method (Russell and Barton, 1992) aligns two structures using dynamic programming, as does the STRUCTAL method (Subbiah *et al.*, 1993), and the Combinatorial Extension (CE) algorithm (Shindyalov and Bourne, 1998). The Sequential Structure Alignment Program (SSAP) (Taylor and Orengo, 1989; Orengo and Taylor, 1996) and CATHEDRAL (Greene *et al.*, 2007) use double dynamic programming. Since SSAP is used in several analyses described in this thesis, the outline of the method will be given in more detail below.

SSAP uses Needleman-Wunsch dynamic programming algorithms, which have otherwise been used for sequence alignment. Two levels of dynamic programming are used. For each structure in a pairwise comparison, a set of vectors connecting a particular β-carbon atom in a protein structure to all other β-carbon atoms is described. Then, these vectors are compared for each pair of potentially equivalent residues between the two proteins, for example, the pairs of residues that have similar ϕ and ψ torsion angles, and accessible areas. The first level of dynamic programming is used at this stage to find the best alignment of the vector sets of these potentially equivalent pairs of residues. It uses a 2D residue level score matrix to score the similarity of each residue pair and determines the best path through

the matrix using dynamic programming. For those pairs of residues with a path score greater than a threshold, the scores from the optimal path are added to a 2D summary score matrix. All potentially equivalent residue pairs are compared in this manner and then the second level of dynamic programming is used to find the best path through the 2D summary score matrix. This defines the aligned residues between the proteins (Sillitoe and Orengo, 2002).

By contrast with the other residue-based methods above, the DALI method (Holm and Sander, 1993) uses simulated annealing to handle indels. A two-stage approach is used to build an alignment of similar hexapeptide backbone fragments between two protein structures (Koehl, 2006).

When assessing the outcome of these methods, i.e. the significance of the similarity between two structures, the quality of the superposition, and the number of residues in the alignment are required. The Root Mean Square Deviation (RMSD) is used as a measure of the superposition quality. It measures the average distance between the alpha-carbons in superposed proteins (see definition in Equation 1.3, from Rao and Rossmann (1973) and Maiorov and Crippen (1994)).

$$D^2_{dis}(A,B) = (n(n-1)/2)^{-1} \sum_{i<j}^{n} (d_{Aij} - d_{Bij})^2 \tag{1.3}$$

d_{Aij} and d_{Bij} represent the corresponding distances between the ith atom and the jth atom.

Whilst studies have suggested a correlation between structural and functional similarity (e.g. Redfern *et al.* (2008)), there are no clear thresholds on RMSD between superposed structures to confirm functional similarity between a pair of proteins. However, when aligning proteins using the SSAP method, a score of 85 out of 100 is usually associated with similarity in function.

Similarly, although classifying a protein by its structural fold can suggest a functional annotation, there are a number of folds that are highly functionally divergent (Martin *et al.*, 1998). Furthermore, as mentioned already, many domains in the CATH database belong to functionally diverse superfamilies. For example, $< 4\%$ of

superfamilies account for $> 40\%$ of all domains in CATH and these large superfamilies are very divergent in structure and function (Dessailly *et al.*, 2009).

In the TIM barrel fold, there are 28 different CATH homologous superfamilies, where the fold supports a different function in each (Lee *et al.*, 2007). Whilst some approaches take into account the functional clues from a protein's overall fold, additional methods that also investigate local structural features are required (discussed below).

1.7.2.2 Surface structural features

The surface of a protein can provide a number of functional clues, arising from the location and size of clefts and pockets. Cofactors, substrates and regulatory elements can bind in surface clefts and most enzymes have their active site in one of the two largest clefts (Laskowski *et al.*, 1996). There are also cases where several clefts are involved in binding interactions, for example in homodimer protein-protein interactions (Laskowski *et al.*, 1996). Laskowski *et al.* (1996) found that in over 83% single-chain enzymes, the ligand binds in the largest cleft and that this cleft is usually much bigger than the others. This makes it relatively easy to identify possible active sites simply by searching the surface for large clefts. When there is no large cleft visible however, bound ligands are used to pinpoint the active site.

It is possible to annotate unknown proteins by comparing these clefts against a library of known active sites. For example, one resource is the electrostatic surface of functional site (eF-Site) database (Kinoshita and Nakamura, 2004). This resource contains electrostatic-potential surface information, which can be compared to an unknown protein model to identify similar regions of charge that are used in binding and interactions.

Electrostatic potential maps on a protein surface are also used to identify DNA-binding regions. As DNA is negatively-charged, the protein surfaces that binds the molecule tend to be the most positively charged on the structure. There are a limited number of structural motifs that are used to bind DNA, with the helix-turn-helix

(HTH) motif being the most popular. As many proteins that do not bind DNA also contain this motif, methods such as eF-Site are used to find true positives (Kinoshita and Nakamura, 2004).

Another example is the pvSOAR (pocket and void Surface of Amino Acid Residues) resource, which compares the protein surfaces of geometrically defined pockets and voids (Binkowski *et al.*, 2005)

1.7.2.3 Using local 3D template methods

Certain proteins, particularly enzymes, have a small number of residues in a localised 3D area that are key to the function. For example, an enzyme's catalytic function is determined by the catalytic residues in its active site. For DNA-binding proteins, there will be a number of residues on the surface that are required for the specificity of binding to DNA. Such examples of highly-conserved conformations can be transformed into 3D residue templates. 3D templates can be constructed manually through expert knowledge, literature searches, or they can be generated automatically through sequence comparisons and structural comparisons (Watson *et al.*, 2005). This type of method can be useful in finding the small changes in a binding or active site that cause functional alterations, which might be missed through a global structure-matching method (Lee *et al.*, 2007).

Other template methods ask the user to define residue patterns to search for. For example, the ASSAM (Spriggs *et al.*, 2003) and the RIGOR/SPASM programs (Kleywegt, 1999), and the MSDsite database (Golovin *et al.*, 2005), now known as PDBeMotif, which all run user-defined pattern queries against the PDB (Watson *et al.*, 2005).

Structural templates have also been created from identified catalytic residues in the CSA, which are compared to structures of unknown function to potentially infer an EC number annotation (Porter *et al.*, 2004). The same group later created a library of structural templates based on the active site areas of non-redundant CSA families (Torrance *et al.*, 2005). This library was then used to see how catalytic

sites vary within families having the same enzyme classification. A representative template was chosen for each family and each family member was superimposed with the representative template. The degree of difference between this pair was calculated by RMSD. Sequence identity between the pair was also used to quantify evolutionary divergence. Catalytic site structure was found to be highly conserved, even where sequence identity was very low. This is thought to be due to the high functional constraints placed upon the catalytic residues, which need to be in an optimal position to be most effective during catalysis (Torrance *et al.*, 2005).

The Comparison of Protein Active Site Structures (CPASS) database (Powers *et al.*, 2006) stores experimentally-identified ligand-binding sites within PDB structures to infer active site residues, biological function and to aid drug discovery. CPASS took the 34,000 protein structures (derived from X-ray crystallography and NMR) in the PDB in 2006 and stored only protein structures with a ligand bound. The CPASS program compares any set of ligand-defined active sites and assesses the sequence and structural similarity, where the sequence does not have to be continuous (Powers *et al.*, 2006).

1.7.2.4 Using a combination of sequence- and structure-based methods

Although it can be helpful to exploit structural information in the identification of protein functions, the quality of the structure and the possibility of artifacts needs to be kept in mind. If non-cognate ligands are used to stabilise the structure for crystallisation, this could cause a conformational change in the structure, changing the native structure. In general it is recommended to use multiple approaches to increase confidence.

Obtaining predictions from multiple sources will increase the likelihood of getting the correct function prediction for a query protein. The ProFunc (Laskowski *et al.*, 2005) and ProKnow (Pal and Eisenberg, 2005) servers combine multiple methods and some of the data sources described above. ProFunc uses BLAST and HMM searches with fold matching, residue template-based, surface-cleft analysis, and sequence

similarity scans. ProKnow performs similar analysis and also uses a probability model to assign GO annotations to the query protein. It also extracts features from the query structure including fold, sequence motifs and structural motifs (Pal and Eisenberg, 2005).

1.7.3 Assessment of protein function prediction

The simplest method for validating protein function prediction methods in enzyme superfamilies is to use EC annotations. One can determine whether the prediction method has annotated each sequence in a superfamily with the correct EC term, using the four hierarchical levels in the EC classification.

Another method of validating protein function prediction is to use the Structure Function Linkage Database (SFLD) (Akiva *et al.*, 2014; Pegg *et al.*, 2006). There are nine manually-curated functionally diverse superfamilies in the SFLD that are referred to as a 'gold standard dataset', and which have been used to validate a number of function prediction methods. Gold standard families consist of either experimentally determined functions or sequences that have a high sequence similarity using BLAST with E-values $< 1\text{x}10^{-175}$ to experimentally characterise sequences (Pegg *et al.*, 2006). Superfamily members are clustered into subgroups based upon sequence similarity information. These subgroups are then further clustered in families, which represent enzymes that catalyse the same reaction using a shared reaction step (Akiva *et al.*, 2014).

A recent community-based protein function validation effort is COMputational BRidges to EXperiments (COMBREX), based in the USA (Anton *et al.*, 2013). Its main aim is to provide a traceable link between protein function predictions made *in silico* and experiments that validate the protein function. First protein function predictions are prioritised according to how confident the predictions are and how many other homologous proteins could be annotated if this protein was experimentally characterised. The enterprise then funds the experiments needed to test the function of the top-ranked predictions. The COMBREX database comprises

approximately 3.3 million proteins from $>$ 1000 complete microbial genome projects that are associated with approximately 2.5 million function predictions. In the first year of the project, 140 proteins were tested experimentally. This number may seem low but a sequence search of these 140 proteins with BLAST and an E-value $\leqslant$ 1x10^{-05} reports sequence similarity to over 60,000 proteins, for which the characterisation could be of use. The 140 proteins also map to 8 domains of unknown function (DUFs) in Pfam (Anton *et al.*, 2013).

The first large-scale community assessment of protein function annotation (CAFA) experiment took place in 2011 and highlighted the issues involved in annotation (Radivojac *et al.*, 2013). Fifty-four methods from across the world were submitted, where each method predicted the protein function of 48,298 target protein sequences from 7 eukaryotic and 11 prokaryotic organisms. As protein function can be described in many ways, the classification schemes from the Gene Ontology (GO) Consortium (Ashburner *et al.*, 2000) were used, which provide a controlled vocabulary to annotate protein function related to: 1) biological process, 2) cellular component, and 3) molecular function.

The methods submitted to the CAFA experiment used a mixture of biological and computational concepts. The majority of methods used sequence alignment information, protein structure, protein-protein interactions or gene expression data. Many methods used machine learning techniques to combine this information and a few methods used literature mining (Radivojac *et al.*, 2013).

The accuracy of each method was assessed using the maximum F-measure score (F_{max}), which considers the precision and the recall to calculate the score. The precision (pr) value is the proportion of the positive predictions that are true positives, as opposed to false positives (see Equation 1.4). The recall (rc) value is the proportion of positive predictions that are correctly identified (see Equation 1.5). The F_{max} score ranges from zero to one, with one being the best score and zero the worst (see Equation 1.6, taken from (Radivojac *et al.*, 2013)). F_{max} was assessed using a number of different chosen thresholds, represented by t.

$$pr = \frac{TP}{TP + FP} \tag{1.4}$$

$$rc = \frac{TP}{TP + FN} \tag{1.5}$$

$$F_{max} = max(t) \left\{ \frac{2 \cdot pr(t) \cdot rc(t)}{pr(t) + rc(t)} \right\} \tag{1.6}$$

All methods were compared against two baseline methods; Basic Local Alignment Search Tool (BLAST) hits (Altschul *et al.*, 1997) and a naive method, where each GO term for each sequence target was scored according to the GO term's abundance in the SwissProt database. All methods were also evaluated against PSI-BLAST predictions, however PSI-BLAST performed no better than BLAST and so only BLAST was reported. When predicting protein functions in the GO category of Molecular Function, the F_{max} for both BLAST and PSI-BLAST was 0.38. For protein function prediction in the GO category of Biological Process, PSI-BLAST reached a F_{max} score of 0.24 and BLAST a score of 0.26. A total of 26 models performed better than BLAST when predicting function in the Biological Process category, whereas 33 models performed better than BLAST in the Molecular Function category.

The ten methods that performed the best overall in the Molecular Function and the Biological Process ontology categories included the use of: gene expression data, machine learning techniques, and sequence homology information. The DFX method, used to generate functional families in CATH, came 7th out of the 56 groups that took part and was the best domain-based method (see Figure 1.3). The top method by the Jones group at UCL in both ontology categories combined predictions based on homology, sequence composition, and gene expression data (Radivojac *et al.*, 2013).

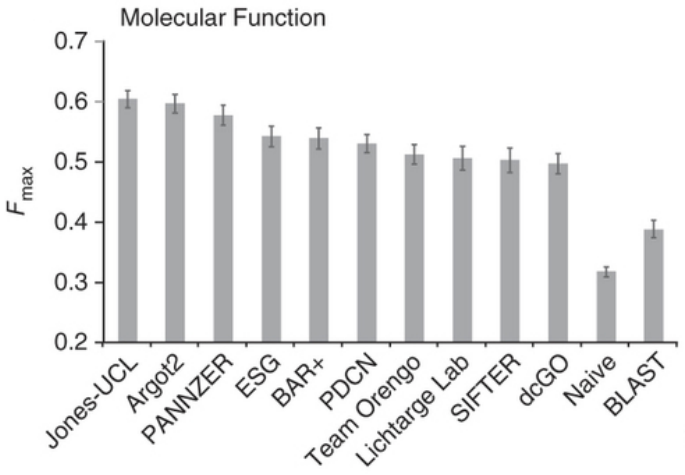


Figure 1.3: Plot to the top performing protein function prediction methods in the first CAFA experiment. Taken from Radivojac *et al.* (2013).

1.7.4 Outline of thesis

The first work chapter in this thesis explores the extent to which different relatives in functionally diverse protein domain superfamilies exploit different functional sites. This was done by examining all known information on functional sites contained in the CSA and IBIS resources. Subsequently, functional sites, identified in CATH functional families, were compared to explore whether there was any preference for different families to use a common functional site.

The second work chapter examines how catalytic machinery can change within enzyme domain superfamilies. A protocol has been developed to compare catalytic residues between functional families and score their similarity in physicochemical properties. Reaction mechanism information has also been used to examine whether a change in the mechanism can be associated with a change in catalytic machinery.

The third and final work chapter discusses the processing and characterisation of metagenome data. The CATH functional families have been used to characterise and compare the functional capabilities of bacteria in the oral and gut metagenome data.

Chapter 2

Identification and Characterisation of Functional Diversity in Protein Domain Superfamilies

2.1 Introduction

2.1.1 Identifying functional site residues

There are a number of manually curated resources providing information on functionally important residues in proteins. For example, the Catalytic Site Atlas (CSA) and the NCBI Inferred Biomolecular Interactions Server (IBIS). These resources report experimentally determined key functional residues involved in catalytic reactions and interaction binding interfaces, respectively. However not all known protein sequences have experimentally determined functional residue annotations. A number of algorithms have therefore been established to predict functional residues.

Detecting functional residues within a superfamily is a difficult task. Many approaches aim to identify conserved residue sites in a multiple sequence alignment on the assumption that functionally important sites are likely to be conserved. However, the multiple sequence alignment (MSA) that one uses to detect these residues must contain relatives that are functionally similar. Some superfamilies however contain a large numbers of relatives that are structurally and functionally diverse. In these cases, subclassification into functional subfamilies can be performed to produce good quality sequence alignments.

However, the majority ($\sim 96\%$) of superfamilies in the CATH database are small and since their members often carry out the same general biological function, the amino acid residues responsible for carrying out this function are expected to be conserved throughout evolution, i.e. they are conserved across all sequences in a

superfamily alignment. For example, the members of the serine protease superfamily are well known for the preservation of their catalytic triad in the active site where 3 residue positions are completely conserved across all members (Drenth *et al.*, 1972; Kraut, 1977).

If one can identify the conserved residue positions in a protein superfamily sequence alignment then it is likely that these positions could play a structural and/or functional role in the protein. Positions with a structural role for example may be important in maintaining the stability of the protein or important in folding. As mentioned already, positions with a functional role may be catalytic residues, or may be involved in interface interactions such as ligand binding, protein-protein interactions, and protein-DNA binding.

2.1.1.1 Identifying conserved residues

Many different types of algorithms have been developed to study the evolution of homologous protein sequences and identify conserved sites, ranging from simple majority fraction (Wu and Kabat, 1970), to entropy or mutual information-based algorithms (Shannon and Weaver, 1949; Wu and Kabat, 1970; Mihalek *et al.*, 2004), and statistical estimations of residue mutability with methods such as Rate4Site (Pupko *et al.*, 2002). By identifying all of the conserved residues, these methods can be used to identify residues that are preserved throughout evolution, for example residues in enzyme active sites.

Methods based on physicochemical properties Livingstone and Barton (1993) developed the program, Analysis of Multiply Aligned Sequences (AMAS), which calculates a conservation score based on the physicochemical properties observed at each sequence position of a MSA. At each MSA position in a pre-defined subfamily, the general physicochemical properties of each amino acid are defined according to Taylor (1986) and Zvelebil *et al.* (1987). Two different methods are subsequently applied, the first (similar to Zvelebil *et al.* (1987)) scores an alignment position as being 'positively' conserved if the amino acids all have the same physicochemical prop-

erties and 'negatively' conserved if the amino acids have different properties. The second method within AMAS only calculates conservation of the positions whose amino acids have the same physicochemical properties (Livingstone and Barton, 1993).

Methods based on entropy Scorecons is another approach, which was developed by Valdar (2002) to quantify the conservation of each residue position in a protein sequence alignment. Each position is assigned an evolutionary conservation score between 0 and 1, where 0 indicates zero conservation at that position, through to 1 where the residue at that position is completely conserved. Amino acid diversity at each position is calculated using amino acid similarity information from a Dayhoff-like mutation data matrix (Jones *et al.*, 1992). The overall score sums up the contributions from each individual sequence, and sequences are weighted inversely with their redundancy in the alignment (Valdar, 2002). This method was reviewed along with 13 other methods in Manning *et al.* (2008) and was reported to consistently score in third place.

Phylogenetic tree-based approaches In addition to Lichtarge *et al.* (1996), a number of other groups also use a phylogenetic tree-based approach. Consurf (Glaser *et al.*, 2003; Ashkenazy *et al.*, 2010; Celniker *et al.*, 2013) calculates the evolutionary rate of each position in an MSA to determine which positions are highly variable throughout evolution (i.e. not conserved), and those that evolve slowly and are therefore highly conserved. The evolutionary rate is calculated through an empirical Bayesian method or a maximum likelihood method, and the phylogenetic tree is used to display the evolutionary relationships between MSA sequences. Conservation scores are mapped onto the 3D structure of a family member to display any clusters of highly conserved residues on the protein surface, which are inferred to be functionally important residues. The MINER algorithm (La and Livesay (2005)) uses evolutionary information from an MSA to calculate the phylogenetic similarity between a local (sliding window) region of the MSA and the whole alignment.

From this, phylogenetic motifs (PMs) are identified, which are shown to structurally cluster around key functional residues.

2.1.2 Identifying specificity determining positions in functionally diverse superfamilies

As conserved residues can relate to either structural or functional roles, algorithms have been developed to distinguish between these two roles. Within a functionally diverse superfamily, functional residues may be conserved throughout the superfamily however the physicochemical property of the residue may change between subfamilies performing different functions, for example they may be associated with changes in substrate specificity. Such residues are described as specificity determining positions (SDPs).

An SDP can be identified as an MSA position that is highly conserved across two subfamilies in the same homologous superfamily but the conserved amino acid residue in each subfamily is different. By examining these different specificities, the degree of functional diversity can be examined within a given protein family. Where functions are known for the different subfamilies, the SDP information can then be used to assign specific functions to query sequences (Capra and Singh, 2008). The first approaches to the identification of SDPs were published in the mid-1990s from the groups of Sander (Casari *et al.* (1995)), who used principal component analysis to identify the functional residues from an MSA of homologues, and Lichtarge *et al.* (1996), who developed the Evolutionary Trace (ET) method, which uses conserved sequence patterns in homologous proteins as well as residue mutation information from the evolutionary relationships to detect changes in function.

Since then, a variety of algorithms have been produced to identify functional residues.

Entropy and mutual information-based algorithms The Funshift database (Abhiman and Sonnhammer (2005a,b)) identifies SDPs using a method described

by Sjölander (1998) where the amino acid distribution of each MSA position is calculated and the relative entropy between two subfamilies computed. The relative entropy across the MSA positions is accumulated and converted into a Z-score, which would be high (above 0.5) if the amino acid distributions between two subfamilies were very different (Abhiman and Sonnhammer (2005a)). Alignment positions that are not conserved between subfamilies and evolved at different evolutionary rates are identified using the maximum likelihood ratio-based (LRT) method by Sjölander (1998) and Knudsen and Miyamoto (2001).

Hannenhalli and Russell (2000) present a method that, given a protein family MSA and a set of protein sequences grouped into subfamilies (e.g. by EC number), identifies SDPs through the comparison of subfamily-specific sequence profiles and also through the analysis of entropy at each alignment position. Alignment positions with significantly high relative entropy were found to correlate with SDP positions known to be involved in defining subfamilies for enzymes including protein kinases and trypsin-like serine proteases.

Mirny and Gelfand (2002) used an entropy and mutual information-based approach to identify SDPs between subgroups of paralogues and orthologues. Orthologous proteins generally have the same biochemical function, however paralogous proteins usually have different specificities as they recognise and bind to different substrates. Orthologous proteins tend to have the same specificity in different organisms as they bind the same substrate. The method first identifies paralogues in a group of homologous proteins and for each of these paralogues, its orthologues are identified in related organisms and an MSA constructed. Mutual information is computed for each alignment position and the likelihood of a position being a SDP is calculated. Kalinina *et al.* (2004) used this method to develop the web-tool SDPpred for users to submit their own MSAs as well as their subfamily definitions to predict SDPs.

2.1.3 Characterising functional site diversity in a large protein domain superfamily

Dessailly *et al.* (2010) performed a detailed analysis of the large, functionally and structurally diverse HUPs domain superfamily. The name "HUP" is derived from different members of the superfamily, which include "High signature" proteins, the "Universal stress protein A", and "PP-ATPase" (Aravind *et al.*, 2002). As mentioned already, the majority of homologous superfamilies have relatively small populations and the relatives perform the same, or very similar function. Less than 4% of superfamilies (accounting for more than 50% of domains in CATH) contain many relatives which are very functionally diverse (Cuff *et al.*, 2011). To understand functional divergence better in these large superfamilies, the members of the HUP domain superfamily were studied in some detail. Relatives were subclassified into functional groups if they performed similar reactions but had different substrate specificities. Functionally important residues were found to be more conserved within these subgroups than between subgroups. Catalytic residues were found to be more conserved than ligand-binding residues. They also found that some functional groups used different protein interfaces to bind with different protein domain partners (Dessailly *et al.*, 2010).

2.1.4 Aims and objectives

This chapter follows on from the analysis of the HUPs superfamily by Dessailly *et al.* (2010) by examining functional site diversity across all functionally diverse CATH superfamilies.

Functionally diverse superfamilies are identified based on the number of EC and GO terms they have and the number of diverse functional families within them.

Variations in functional site residue locations across a superfamily are identified by mapping all known sites for a superfamily onto a representative structure.

Two different methods are used for identifying the functional families within a

superfamily. The functional purity of the families produced by these methods are assessed and the most coherent functional family dataset was then used to identify common and unique functional sites within a superfamily.

Finally, examples of superfamilies with highly conserved and highly variable functional sites will be discussed.

2.2 Methods

In this work, data from CATH (version 3.5) and Gene3D (version 11.0) (Cuff *et al.*, 2011) have been used to analyse functional diversity within CATH superfamilies, and the extent to which superfamily relatives share the same functional site residues. The first part of the work involved mapping all of the functional site data available for a given superfamily onto a structural representative for that superfamily. This part of the project was done in collaboration with Benoit Dessailly, who had been involved in the initial design of the project before taking up a Fellowship in Japan.

As well as examining the variation in location of functional sites across domain superfamilies, we wanted to examine whether some sites were commonly used by relatives having different functions. For this analysis we used a functional subclassification of CATH (functional families, or FunFams). FunFams were generated by a small team of researchers in the Orengo Group. My contribution was to assess the functional and structural purity (also referred to as coherence) of different functional subclassifications.

Work from this chapter has been published in Dessailly, Dawson, Mizuguchi, and Orengo (2013).

2.2.1 Domain superfamily data

CATH functional families Functional families were introduced to the CATH-Gene3D classification in 2012 (Lees *et al.*, 2012) using a two-step approach. First, all superfamily sequences are compared to generate a hierarchical tree. Second, similarities between sequences at each node of the tree are analysed to determine where to cut the tree to produce functionally coherent sequence clusters.

In the first step, the GeMMA algorithm (Lee *et al.*, 2010) clusters relatives in a superfamily using a hierarchical agglomerative clustering algorithm to produce a tree of clusters built from the leaf nodes to the root node. This iterative approach first clusters close homologues, i.e. sequences with at least 90% sequence identity, using

the program CD-HIT (Li and Godzik, 2006). For each of these clusters, MSAs are constructed using MAFFT (Katoh *et al.*, 2005). In the second iteration, clusters are compared using a profile-profile approach that exploits the COmparison of Multiple Protein Alignments with Assessment of Statistical Significance (COMPASS) set of tools (Sadreyev and Grishin, 2003). COMPASS takes two MSAs as input and from these builds two position-specific scoring matrix (PSSM) profiles for comparison purposes and reports the similarity of profile pairs. GeMMA then identifies the profiles with the highest similarity which are then merged. This continues until one cluster remains. The results of clustering can be visualised as a tree of clusters for the superfamily.

The final tree of clusters is partitioned by cutting the tree. In the original 2012 approach used to generate functional families, the nodes are merged if their similarity value has an E-value less than $1x10^{-10}$, otherwise they are kept separate. This approach is referred to as the 'unsupervised', or Funfamer, method (Lee *et al.*, 2010). Functional families generated by this method will be described as FunFam_{SEQ} functional families (Figure A.1). The FunFam_{SEQ} functional families have been previously benchmarked against the Structure Function Linkage Database (SFLD) (Pegg *et al.*, 2006).

A more recent approach to generating the functional families was developed using a 'supervised' protocol (Rentzsch and Orengo, 2012). This approach (DFX) detects and accounts for functional 'chaining' within the tree of clusters. 'Chaining' refers to instances of protein domain sequences in a superfamily that cluster in an unexpected way. In DFX, GO annotation data is used to ensure functional coherence in each functional family, and clusters are only merged if they contain coherent GO terms. However, in some superfamilies the sequence similarity reflected in the COMPASS scores appears to contradict GO term similarity so that domain relatives apparently having different functions are preferentially merged in the hierarchy. This phenomenon usually arises because in these superfamilies, domains have a generic functional role that remains unchanged despite the different functional contexts

(reflected in different GO terms for their parent proteins) in which the relatives appear. The DFX method is described in more detail in Rentzsch and Orengo (2012) and the functional families produced by this method will be referred to as FunFam$_{GO}$s (Figure A.1).

FunFam$_{GO}$s were shown to perform well in predicting functions for uncharacterised sequences. The DFX method came 7th out of 56 groups in the CAFA independent assessment of function prediction (Radivojac *et al.*, 2013).

Characterising superfamily diversity by the number of s60 clusters "s60 clusters" are groups of domain sequences clustered at 60% sequence identity (Figure A.1). Whereas some superfamily relatives cannot be grouped into functional families because they lack GO functional terms or they do not match the HMMs generated for a given functional family, all relatives can be clustered into s60s. The number of s60 clusters was used to provide a measure of superfamily diversity. This analysis used the 1,456 superfamilies that have two or more s60 clusters.

2.2.2 Assessment of functional purity within functional families

Functional families are generated as clusters of sequences that code for protein domains carrying out a highly similar, if not the same, biological function. The same functional residues are expected to be present in the same sequence position throughout, i.e. the functional residues are expected to be completely conserved.

We first assessed and compared the functional purity of FunFam$_{SEQ}$ and FunFam$_{GO}$ functional families to determine which were most functionally coherent. These would then be used for the analysis of common functional sites. The FunFams were assessed by examining whether conserved residues, i.e. a position in an MSA where the amino acid residues are the same throughout all relatives, were enriched in known catalytic and binding functional residues. Catalytic residues with literature-based evidence were taken from the Catalytic Site Atlas (CSA) (Porter *et al.*, 2004). Residues

involved in binding interfaces were taken from the Inferred Biological Interactions Server (IBIS) resource (Shoemaker *et al.*, 2012). The binding interfaces included ligand information for protein-protein, protein-nucleic acid, and protein-small ligand interactions. Small ligands encompass the small organic compounds, peptides and ions. All available literature-based functional site information was used for all residues found in structural domains and the data was not limited to either buried or surface residues.

Residues in a multiple alignment position were considered conserved if assigned a score of least 0.7 by the in-house Scorecons program (Valdar, 2002) (described in Introduction, page 53). This value has been previously been determined as identifying a high proportion of known catalytic residues from the CSA (Dessailly *et al.*, 2010).

Enrichment tests were adapted from Rausell *et al.* (2010) and performed to find out whether conserved residues in functional family alignments were significantly enriched in known functional residues. The enrichment (E) value for each functional family was calculated as the proportion of conserved residues that were functional (Pc), minus the proportion of all residues in the protein domains that were functional (Pa) (see Equation 2.1).

$$E = Pc - Pa \tag{2.1}$$

E values were averaged for each superfamily. An unpaired, one-sided Wilcoxon Rank-Sum test (Kruskal, 1957) was run on all averaged enrichment values using the wilcox.test function in R (R Core Team, 2012). This test assessed a P-value for the null hypothesis that the median enrichment value was zero. An enrichment value of zero indicates that the proportion of unconserved functional residues is the same of the proportion of conserved functional residues. The alternative hypothesis assumed that the median enrichment value was greater than zero, i.e. a positive E value, which reflected a greater proportion of conserved functional residues in comparison to unconserved functional residues.

A functional family may contain sequences that do not always represent a complete protein domain, i.e. sequence fragments. This is due to a number of reasons: 1) the correct start and/or end gene positions may not have been correctly defined in UniProt, 2) there are only partial annotations for some genes in a number of completed genomes, or 3) there is no close homologue available to define a gene. Due to such issues there are sometimes protein domain fragments within the domain sequence clusters. These sequence fragments could affect the quality of the functional family MSA, e.g. there may be many gaps caused by aligning the fragments to full-length protein domain sequences. To overcome this issue, sequence fragments were removed from functional family data. Any sequences less than 80% of the average family sequence length were removed.

Once the sequence fragments had been removed, functional family sequences were re-aligned with MAFFT (Katoh and Toh, 2008) and residue conservation scores were calculated with Scorecons (Valdar, 2002). Enrichment tests were then performed for these data as previously described.

2.2.3 Assessment of structural coherence in functional families

To assess the structural coherence of a set of relatives, the relatives were pairwise compared using SSAP (Taylor and Orengo, 1989) and the mean RMSD calculated. Since running SSAP is computationally expensive, structural coherence was only measured for each FunFam$_{SEQ}$, FunFam$_{GO}$, and superfamily within the top 50 most structurally diverse superfamilies. We assumed that if the functional family relatives were found to have high structural similarity in these 50 superfamilies, then functional families in all other less structurally diverse superfamilies would also be structurally coherent. A structurally diverse superfamily here refers to a superfamily with at least five structurally similar groups (SSGs). An SSG is a group of relatives that can be pairwise superposed with each other within a given RMSD threshold. Here we use a cutoff of 5 Å to produce 'SSG5s'.

An unpaired, two-sided Wilcoxon Rank-Sum test was used to determine whether there was a significant difference between the means in the distribution of average RMSD values in the FunFam$_{SEQ}$s and FunFam$_{GO}$s.

2.2.4 Protocol to identify functional site coverage across each superfamily

We first wanted to examine the number of different functional sites (i.e. the coverage of the structure by functional sites) used by relatives across each functionally diverse superfamily using all known functional site data for that superfamily.

Functional site residues were extracted for superfamily members from the CSA (Porter *et al.*, 2004) and IBIS (Shoemaker *et al.*, 2012) resources.

Pairs of superfamily members were aligned using the SSAP algorithm (Taylor and Orengo, 1989) in an all-against-all protocol. The SSAP score calculated from the pairwise comparison was assigned to each domain in the pair. For each domain, all of its assigned SSAP scores were summed and the domain with the highest cumulative SSAP score was chosen as the superfamily representative. The domains with functional site residue information were aligned to the superfamily representative and the functional site residue information was mapped onto the equivalent residues in the superfamily representative.

Functional site coverage for each superfamily representative was calculated as the number of functional site positions mapped onto a representative sequence, divided by the total number of residue positions in the representative.

2.2.5 Protocol to identify common functional sites across each superfamily

Following the examination of functional site coverage, we then wanted to explore whether there were any common sites in the superfamily i.e. sites used by more than one FunFam.

Figure 2.1 describes the steps taken in the identification of common functional sites. Each of these stages will now be discussed.

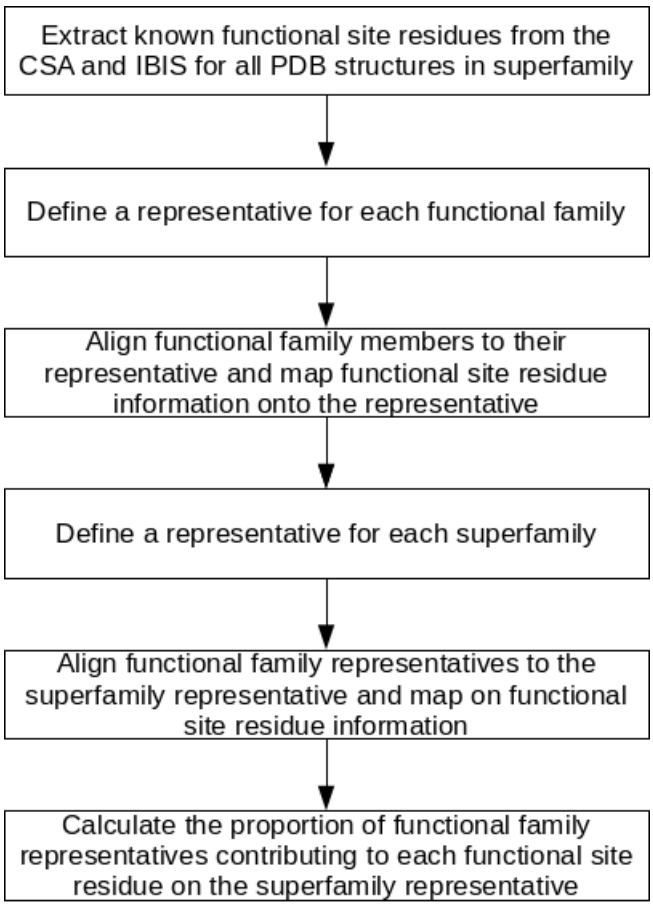


Figure 2.1: Flowchart showing the steps taken to identify common and unique functional sites.

Extracting functional residue site information Functional residues were extracted from the CSA and IBIS for all FunFam$_{SEQ}$ functional family structural domains as described in Section 2.2.4.

Identifying functional family representatives and mapping functional site information onto this representative Pairs of functional family members were aligned using the SSAP algorithm (Taylor and Orengo, 1989) in an all-against-all protocol. The domain with the highest cumulative SSAP score was chosen as the functional family representative. The members with functional site residues were aligned to the functional family representative and the functional site residue infor-

mation mapped onto the equivalent residues in the functional family representative (Figure 2.2). This pairwise mapping was done using an in-house program previously written by Benoit Dessailly for his analysis of the HUPs superfamily.

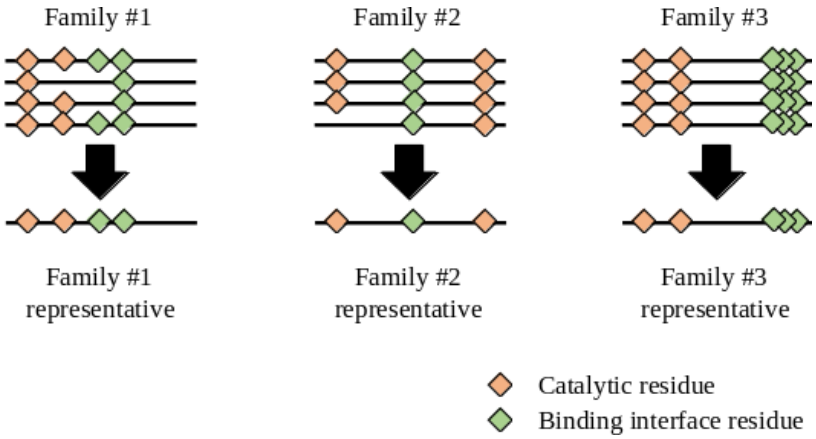


Figure 2.2: Examples of structural domains annotated with functional site residue information from the CSA (in orange) and IBIS (in green). Functional residue information was then inherited onto the functional family representative.

Identifying superfamily representatives and mapping functional site information onto this representative Pairs of functional family representatives were aligned using the SSAP algorithm (Taylor and Orengo, 1989) in an all-against-all protocol. The domain with the highest cumulative SSAP score was chosen as the superfamily representative.

For each functional family, functional site data of CSA and IBIS residues were inherited from each family representative domain onto the superfamily representative domain (Figure 2.3).

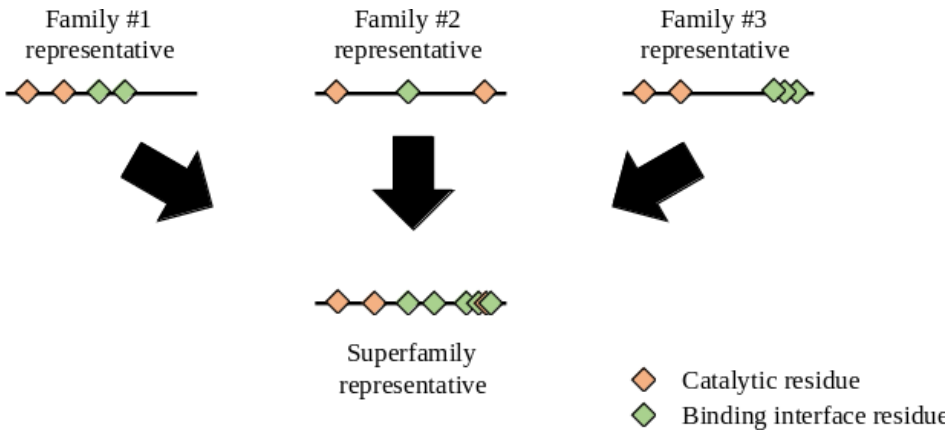


Figure 2.3: Functional residue information was inherited onto the superfamily representative from functional family representatives. Examples of functional site residues from the CSA are shown as orange diamonds and from IBIS as green diamonds.

Identifying common functional sites and characterising functional site diversity The proportion of functional family representatives contributing to each mapped functional site residue on the superfamily representative was calculated. This protocol was used to identify the sites common to functionally diverse relatives, i.e. in FunFams, and to get an idea of the distribution of common and unique sites across the representative domain structure.

Figure 2.4 illustrates the accumulation of catalytic functional site residues from functional families representatives onto the superfamily representative structure. A colour-coding system was used to indicate the proportion of the functional family representatives mapped to each functional site residue on the superfamily representative. For example, if only one of the three example functional families contributed to a mapped position on the superfamily representative, the residue was coloured green (bottom-right structural domain). Meanwhile, if all three functional families within a superfamily mapped to the same functional site residue on the superfamily representative, this residue was coloured red.

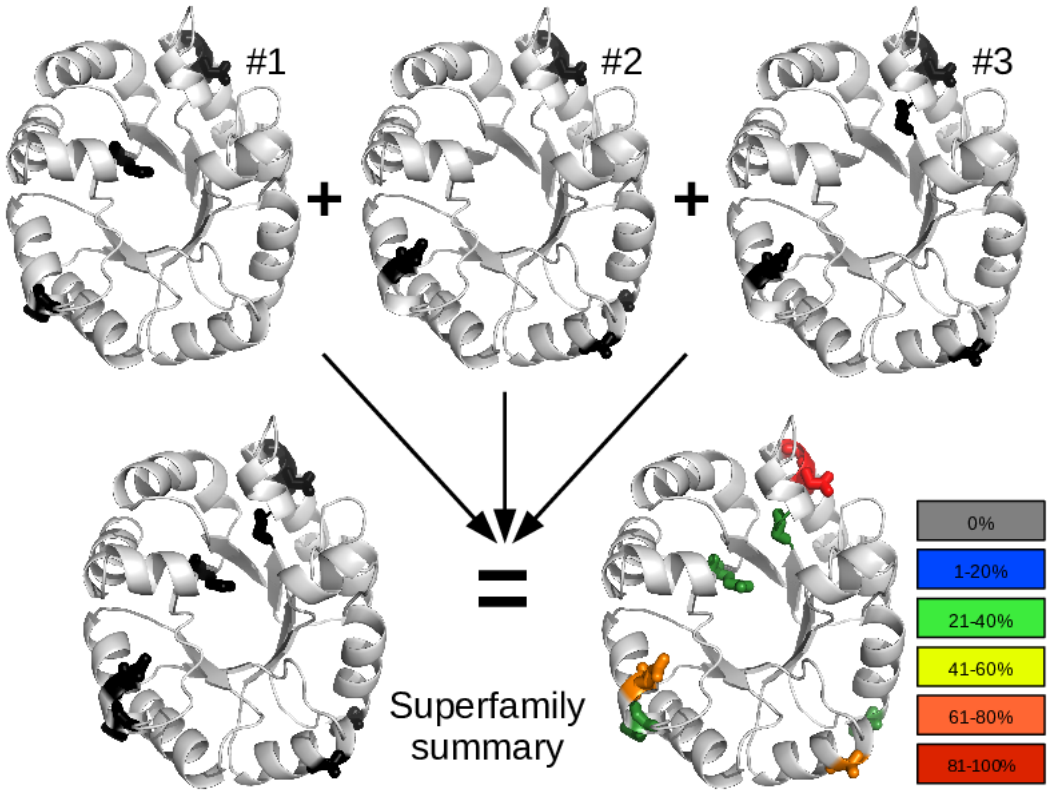


Figure 2.4: Catalytic functional residues mapped onto functional family and superfamily representatives. The top three domains show catalytic functional residues (in black) mapped onto functional family representative domains. These sites were then accumulated onto the superfamily representative (in black, bottom-left structural domain). Each site on the superfamily representative structure was colour-coded to indicate the proportion of functional families with sites that map to that position (bottom-left structural domain).

2.2.6 Post-processing of functional families for the CATH website

A pipeline was written for the processing of all functional families in all CATH superfamilies to make the functional family information available online. The functional families have been available on the CATH website since version 3.5, which was established in 2012 (Sillitoe *et al.*, 2013). For each functional family, the information available included the representative structure with mapped functional site residues highlighted, and the MSA of the family with conserved position marked.

2.3 Results

Domains in some superfamilies are very small and the sites take up a large proportion of the domain. To overcome these issues, superfamilies were only included in the analysis where: 1) the superfamily representative domain was at least 100 amino acids in length, 2) each contributing member of functional site data had a functional coverage of less than 50%. Superfamilies were also only included in the analysis if the superfamily representative had a functional site coverage value above zero.

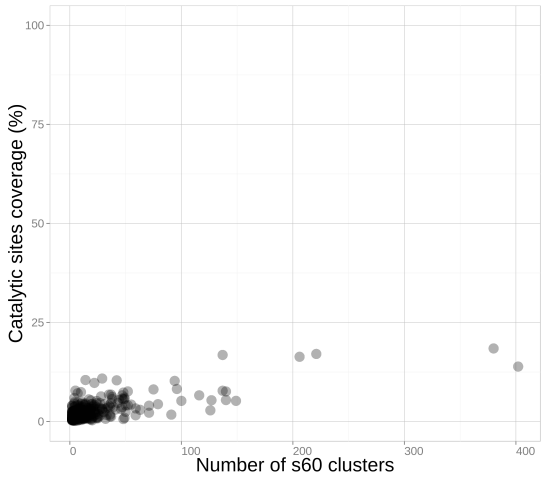
2.3.1 Exploring functional site coverage

Table 2.1 shows the number of superfamilies used in, and excluded from, the analysis after the two filters described above were applied and the superfamily representative domains with no functional site coverage were identified.

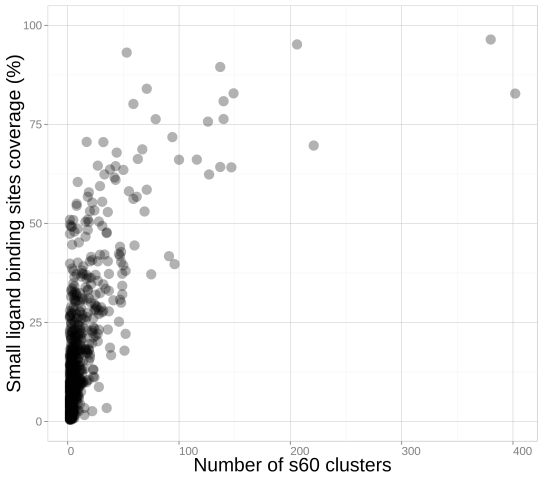
Site Type	# Superfamilies used	# Superfamilies filtered out	# Superfamilies with no coverage
Catalytic Sites	328	53	1,075
Protein-Protein Interfaces	645	433	378
Nucleic Acid Binding Sites	116	107	1,233
Small Ligand Binding Sites	659	274	523

Table 2.1: Number of superfamilies considered in, and excluded from, the dataset for each type of functional site, after applying the two filters and in light of the number of superfamily representative domains with zero functional site coverage.

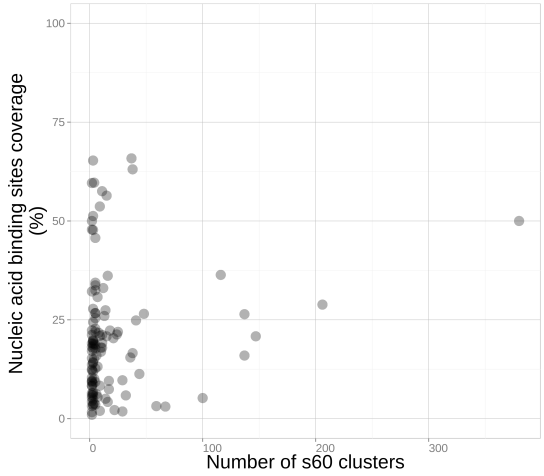
Figure 2.5 shows the functional site coverage calculated for each superfamily analysed, plotted against the diversity of the superfamily measured by the number of s60 clusters in the superfamily. The coverage of the superfamily representative by functional sites is a simple measure of functional site diversity. Superfamilies with a high coverage of functional sites will be those in which the functional sites of relatives occur at many different structural locations.



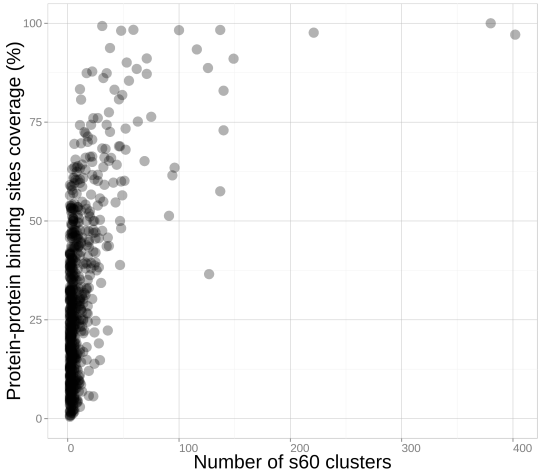
(a) Catalytic Sites



(b) Small Ligand Binding Sites



(c) Nucleic Acid Binding Sites



(d) Protein-Protein Binding Sites

Figure 2.5: Functional site coverage and sequence diversity of domain superfamilies. Each plot shows the data for a specific type of functional site. Each superfamily is represented as a dot in these plots. Each dot is partially transparent to enable visualisation of the data points overlapping. Functional site coverage on the Y-axis is measured as the proportion of residues in the representative that map to a functional site in superfamily members. Superfamily diversity on the X-axis is measured as the number of sequence clusters grouped at 60% sequence identity within each superfamily.

Catalytic residue sites are observed to have the lowest functional site coverage (Figure 2.5a), where at most only 20% of superfamily representative positions have catalytic residues mapped from different family members. Small ligand and protein-protein binding sites (Figures 2.5b and 2.5d) are seen to have the highest levels of functional site coverage. In general, we observe that the more diverse superfamilies frequently have the larger coverage values.

The restriction in catalytic site location could be explained by the relatively small size of an active site, by the limited number of catalytic residues that can be involved in a given catalytic reaction, and by the preference for the same site by superfamily relatives.

For nucleic acid binding sites a maximum of $\sim$ 65% coverage is observed, with most of the superfamilies having a coverage of $\leqslant$ 30% (see Figure 2.5c). This maximum coverage is higher than the maximum coverage observed for catalytic sites. Many superfamilies do not have the ability to bind either DNA or RNA, reflected by a large number of superfamilies (1233 out of 1456, $\sim$ 85%) having a representative coverage of zero for this functional type (see Table 2.1).

For small ligand binding sites, a small number of superfamilies have coverage values over 80% but the majority of superfamilies have less than 40% coverage (see Figure 2.5b). As the ligands being bound are small, fewer residues are required for binding than are needed for binding macromolecules. It may be that high coverage in some of these superfamilies is an artifact of the protein domain binding to different artificial ligands prior to crystallisation and therefore the solved structure contains these bound ligands.

Protein-protein interface binding sites have the highest levels of functional site coverage with up to 100% coverage. For the majority of diverse superfamilies (i.e. those with at least 20 s60 clusters), more than 60% of residue sites on the superfamily representative are mapped by protein-protein interface residues from different superfamily members (See Figure 2.5d).

2.3.2 Assessment of functional purity in functional families

For the next part of the study we wanted to explore the proportion of functional families mapping to a particular site on the superfamily representative. For this we needed to use functionally coherent families. We therefore compared the functional purity of the two families, FunFam$_{SEQ}$ and FunFam$_{GO}$, described in Methods Section 2.2.1.

Functional families ideally represent groups of homologous sequences that carry out related biological functions and for this reason the relatives within them are expected to consist of highly similar or identical functional residues, in both spatial and physicochemical terms.

Residue enrichment analyses were used to assess the functional purity of the FunFam$_{SEQ}$ and FunFam$_{GO}$ functional families generated using the FunFamer and DFX protocols, respectively. Enrichment scores were calculated using both CSA and IBIS functional site data. Figure 2.6 shows that more FunFam$_{SEQ}$s than FunFam$_{GO}$s have a higher percentage of conserved catalytic residues, before and after the removal of partial domain sequences from family sequence alignments.

In the removal of partial domain sequences, also referred to as fragments, a total of 393,430 sequences (out of 4,149,361, 9.5%) were removed from the FunFam$_{SEQ}$s and 521,349 sequences (out of 4,113,867, 12.7%) were removed from the FunFam$_{GO}$s. Following fragment removal there are still approximately 100 FunFams with no conserved catalytic residue positions.

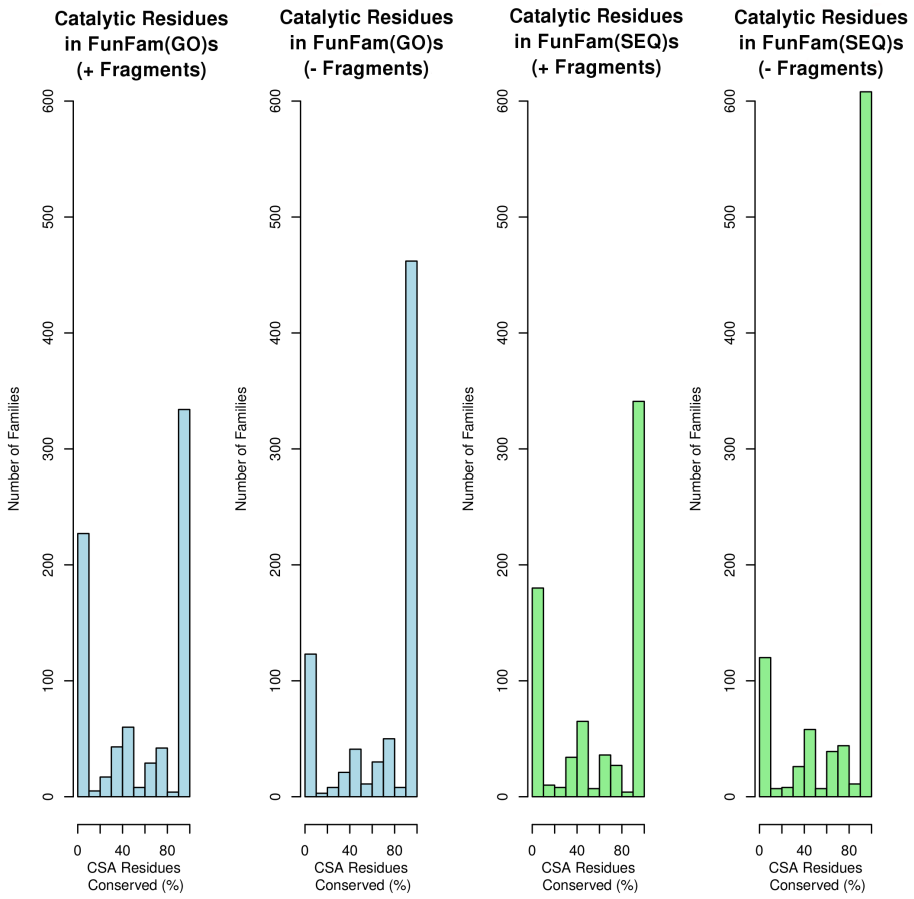


Figure 2.6: The number of functional families with conserved catalytic residues, before and after the removal of partial domain sequences.

The proportion of conserved IBIS functional residues is also higher in FunFam$_{SEQ}$s than FunFam$_{GO}$s, before and after partial sequence fragment removal (see Table 2.2). The proportions are however lower for IBIS residues than for conserved catalytic residues. This is because catalytic residues are more highly conserved and are restricted in their spatial position to a greater degree, as previously shown in Figure 2.5a. Regardless of whether FunFam$_{SEQ}$s or FunFam$_{GO}$s are analysed, the removal of partial sequence fragments shows an increase in the proportion of conserved functional residues, indicating that the family alignment quality has been increased.

Family Type	CSA	IBIS Nucleic Acid Binding	IBIS Protein-Protein Interface	IBIS Small Ligand Binding
FunFam$_{GO}$s (+ fragments)	57.82%	26.15%	21.50%	31.63%
FunFam$_{GO}$s (- fragments)	76.05%	37.32%	33.76%	47.41%
FunFam$_{SEQ}$s (+ fragments)	58.43%	34.70%	24.34%	35.02%
FunFam$_{SEQ}$s (- fragments)	78.33%	47.73%	39.28%	52.03%

Table 2.2: The proportion of CSA and IBIS functional residues that are conserved.

Using the residue enrichment analysis described in Methods, we examined whether the FunFam$_{GO}$ and FunFam$_{SEQ}$ alignments contained conserved residues that were significantly enriched in known functional residues. Wilcoxon Rank-Sum tests reported significant p-values for all functional site types, except for the IBIS protein-protein interface residues which are only significantly enriched in the FunFam$_{SEQ}$ families (Table 2.3).

Family Type	CSA	IBIS Nucleic Acid Binding	IBIS Protein-Protein Interface	IBIS Small Ligand Binding
FunFam$_{GO}$s (+ fragments)	$<2.2 \times 10^{-16}$	5.9×10^{-06}	0.55	$<2.2 \times 10^{-16}$
FunFam$_{GO}$s (- fragments)	$<2.2 \times 10^{-16}$	3.6×10^{-07}	0.35	$<2.2 \times 10^{-16}$
FunFam$_{SEQ}$s (+ fragments)	$<2.2 \times 10^{-16}$	7.0×10^{-09}	0.0095	$<2.2 \times 10^{-16}$
FunFam$_{SEQ}$s (- fragments)	$<2.2 \times 10^{-16}$	2.2×10^{-06}	0.00068	$<2.2 \times 10^{-16}$

Table 2.3: P-values reported from Wilcoxon Rank-Sum tests performed on FunFam$_{GO}$ and FunFam$_{SEQ}$ functional family alignments to observe whether conserved residues were enriched in functional residues.

2.3.3 Assessment of structural coherence in functional families

Figure 2.7 shows that FunFam$_{SEQ}$ functional family relatives superpose with lower RMSD values than relatives within FunFam$_{GO}$s and relatives within superfamilies. As analyses have shown that relatives with significant similarities in structure are more likely to have similar functional roles and protein interactions (Redfern *et al.*, 2008), the FunFam$_{SEQ}$s were considered to be more functionally coherent on the basis of their structural coherence, and therefore selected for further analysis.

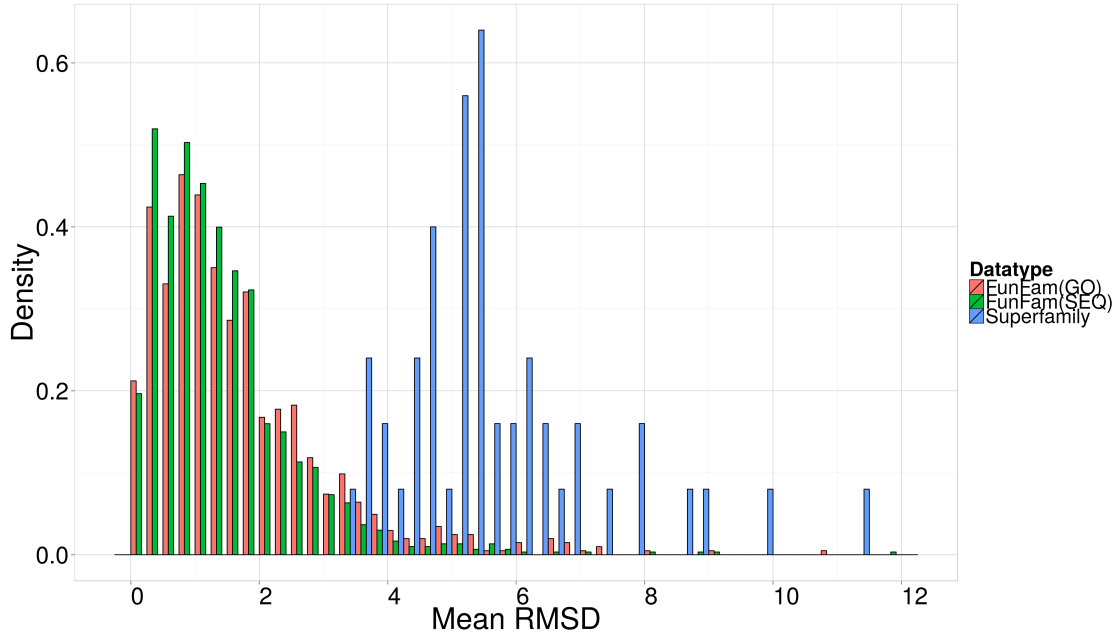


Figure 2.7: Mean RMSD distributions of the top 50 most structurally diverse CATH superfamilies, and the FunFam$_{GO}$s and FunFam$_{SEQ}$s within these superfamilies. Using a Wilcoxon Rank-Sum test, the FunFam$_{GO}$ and FunFam$_{SEQ}$ distributions were found to be significantly different with a p-value < 0.0002253.

2.3.4 Identifying common functional sites in superfamilies

In the previous sections we showed that functional site coverage can be quite high in some superfamilies, i.e. a large proportion of residues are functional site residues, and also for some types of functional sites, e.g. protein-protein interfaces. Here, we examine whether some functional sites on a domain are more frequently used than

others. To do this we examined the proportion of functional families mapped to each functional site residue on the superfamily representative and we identify sites common to multiple functional families. This analysis also revealed sites unique to particular functional families.

An example of a superfamily with high sequence diversity and diverse use of functional sites is shown in Figure 2.8. This figure illustrates the Aldolase Class I superfamily (CATH code 3.20.20.70), whose functional families use different catalytic residue sites. In total, 34 catalytic functional site residues have been mapped onto the superfamily representative.

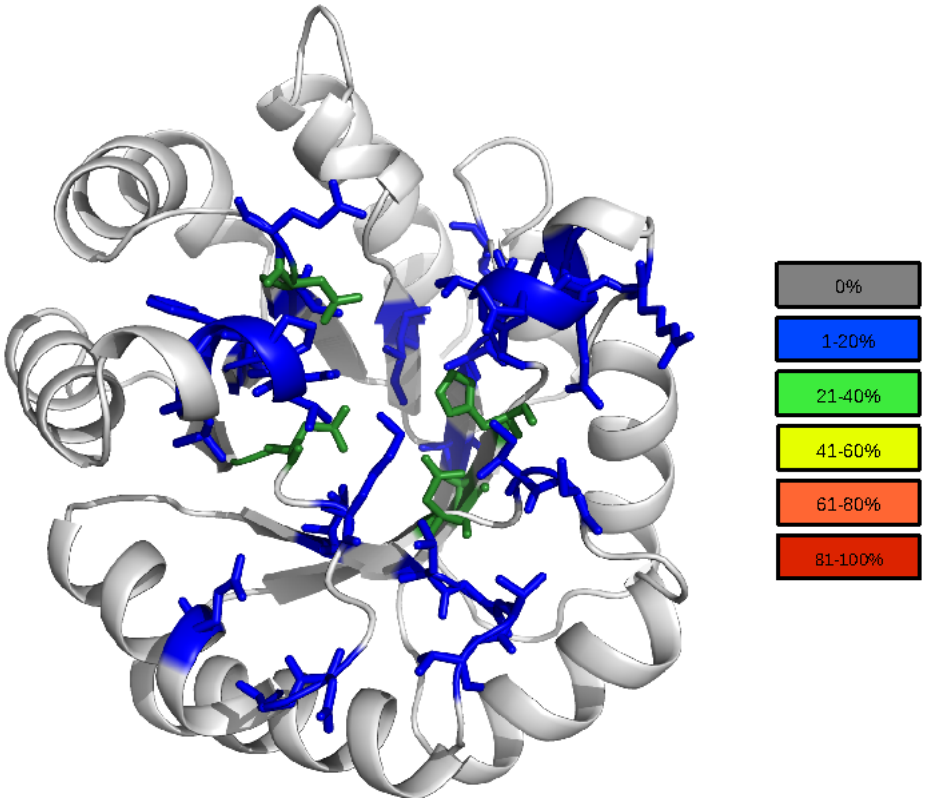


Figure 2.8: Catalytic functional site residues mapped onto the superfamily representative domain 3f4wA00 for the Aldolase Class I superfamily (CATH code 3.20.20.70). Mapped residues are colour-coded according to the proportion of functional families contributing catalytic site data to a given site.

To ensure that our functional family dataset gave similar insights into functional site coverage, we repeated the previous study of functional site coverage as reported in Section 2.3.1. Figure 2.9 shows the functional site coverage for each superfamily, against the functional diversity of the superfamily measured by the number of

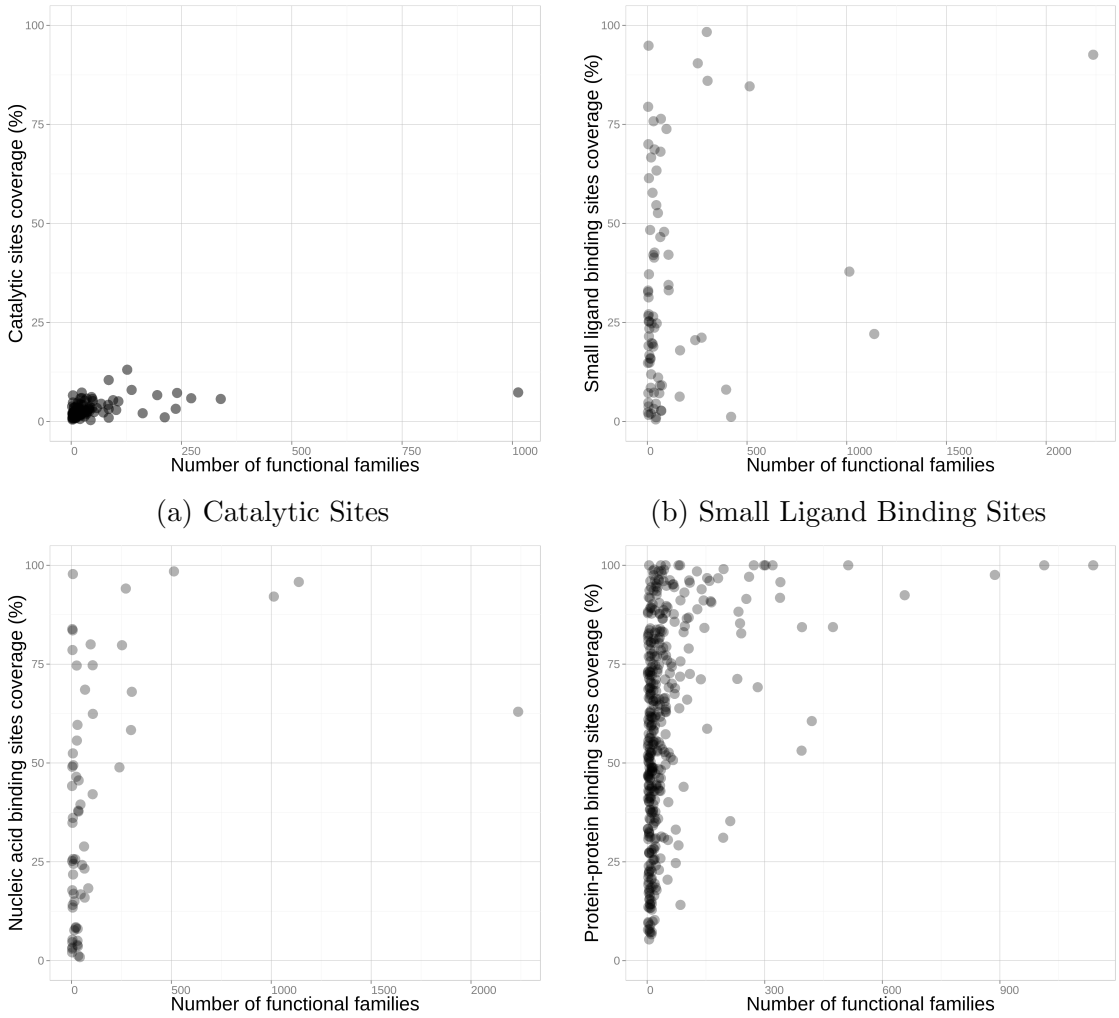
FunFam$_{SEQ}$ functional families.

As seen previously in Figure 2.5a, catalytic residue sites have the lowest superfamily representative coverage values. In the previous analysis, a maximum coverage of 20% was observed and now a similarly low maximum coverage of 13% is reported (see Figure 2.9a). The coverage values calculated from nucleic acid binding site data still show more than half of the superfamilies have $\leqslant$ 30% coverage.

Small ligand binding site data shows coverage values that are below 25% for half of the superfamilies, which is slightly lower than the previous results that showed a large proportion of the superfamilies had less than 40% coverage.

The lower coverage values for catalytic and small ligand binding sites may be due to the fact that FunFams must have at least one experimentally characterised GO term. Therefore not all CATH domain structures with site data are included in the FunFam dataset, i.e. some of the PDB structures are not mapped to FunFams.

Finally, protein-protein interface data also shows the highest levels of superfamily representative coverage, as previously in Figure 2.5d. Again, the majority of the superfamilies have at least 60% of the representative residues mapped to protein-protein interface residues (see Figure 2.9d).



(a) Catalytic Sites

(b) Small Ligand Binding Sites

(c) Nucleic Acid Binding Sites

(d) Protein-Protein Binding Sites

Figure 2.9: Functional site coverage and sequence diversity of domain superfamilies. Each plot shows the data for a specific type of functional site. Each superfamily is represented as a black dot in these plots. Each dot is partially transparent to enable visualisation of the data points overlapping. Functional site coverage on the Y-axis is measured as the proportion of residues in the representative that map to a functional site in one or more superfamily members. Superfamily diversity on the X-axis is measured as the number of FunFam$_{SEQ}$ functional families within each superfamily.

Whilst protein-protein interface binding sites are found in numerous locations across the structure for many superfamilies in the data set, we decided to examine whether there was a preference for a particular site within each superfamily. The red box in Figure 2.10 shows that in 364 (out of 463, $\sim$ 79%) of superfamilies, there is one site (the most common protein-protein interface functional site residue) which is mapped to by 80-100% of their functionally diverse functional families.

The remaining 99 (out of 463, $\sim$ 21%) of superfamilies have common sites which are mapped to by 40-79% of their functionally diverse functional families. This shows that whilst multiple sites can be used, there are preferential sites for interfaces within some superfamilies.

The large coverage observed for protein interfaces is likely to be due to differences in domain and protein partners for functionally different relatives. This has been suggested by previous studies (Todd *et al.*, 2001; Dessailly *et al.*, 2010; Reid *et al.*, 2010). The fact that we observe some overlap between the sites suggests that there may be a part of the surface that has features more suited to forming an interface and that is therefore more frequently exploited.

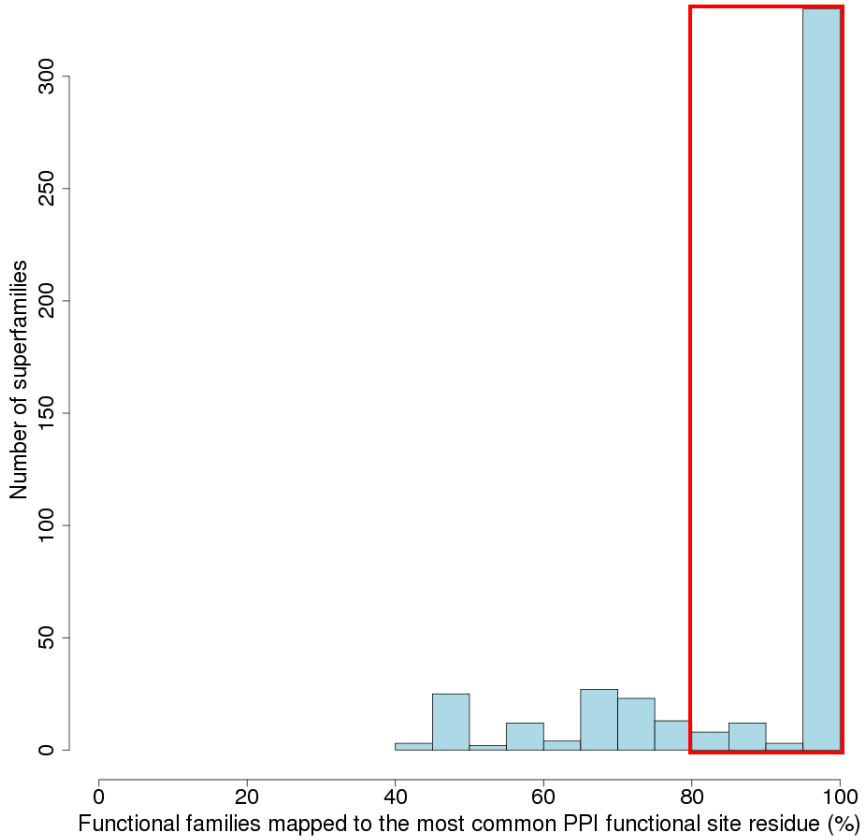


Figure 2.10: The number of superfamilies against the percentage of their functional families mapping to the most common protein-protein interface functional site residue. The red box highlights the 364 ($\sim$ 79%) superfamilies each with 80-100% of their functional families contributing to the most commonly mapped PPI functional site residue.

Some further examples of superfamilies with high or low functional site coverage for protein-protein interfaces and high or low sequence diversity are discussed below.

2.3.4.1 Selected examples

The coverage plots obtained using either s60 or FunFam diversity have shown the range of coverages observed for different superfamilies. Some examples will now be presented from different parts of the coverage plot for interfaces to illustrate some of the varying phenomena observed.

Example of a functionally coherent superfamily with common functional sites The bacterial GTPase-activating protein (GAP) superfamily (CATH code 1.20.120.260) has low functional site coverage (20%) of protein-protein interface binding sites and low sequence diversity with only 3 s60 clusters for 111 bacterial species (see red dot in the bottom left-hand corner of Figure 2.11, which is adapted from Figure 2.5d).

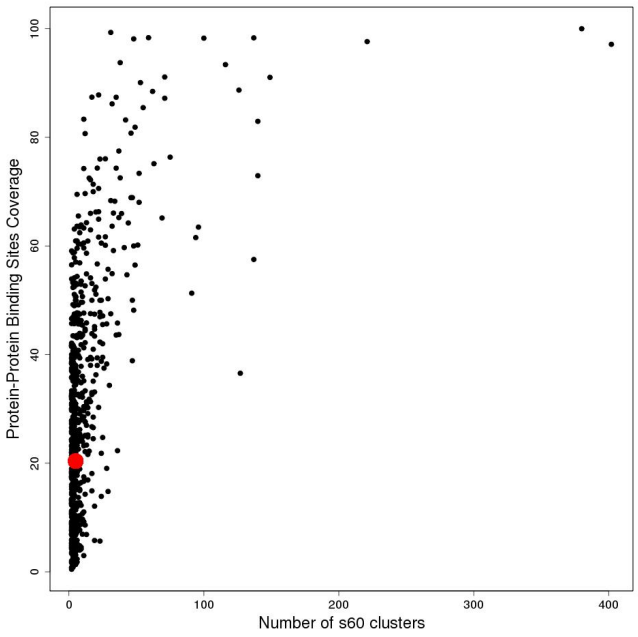
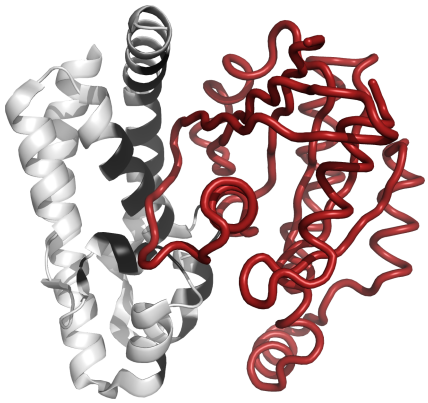


Figure 2.11: The bacterial GAP domain superfamily (CATH code 1.20.120.260) is highlighted in red to show its position amongst all superfamilies in terms of functional site coverage and sequence diversity. This superfamily has low functional site coverage and low sequence diversity. The position of the superfamily in the plot is highlighted in red.

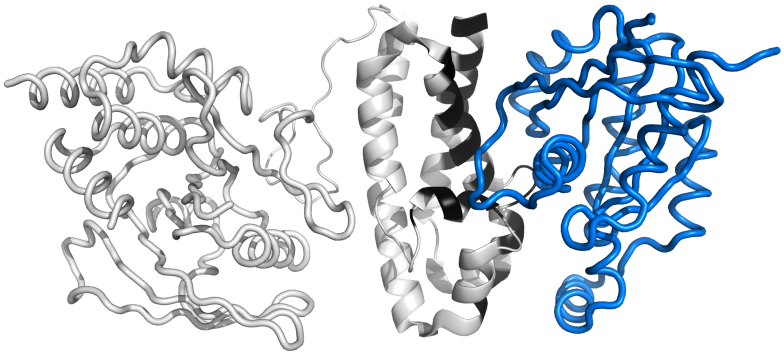
Bacterial type III secretion systems inject GAPs into an eukaryotic host cell where host cell's Rho-family GTPases are targeted to disrupt their maintenance of the host cytoskeleton (Litvak and Selinger, 2003).

The protein-protein interface shown in Figures 2.12a and 2.12b is the same in all relatives. The two relatives illustrated are from different species and have less than 35% sequence identity however they use the same interface binding residues. This similarity in interface is likely to be due to relatives binding to the same protein partner. For example, both relatives bind the human Rac protein (see Figures 2.12a and 2.12b). Litvak and Selinger (2003) constructed a multiple sequence alignment of ten bacterial GAPs and discovered two short highly-conserved motifs across the different species. These two motifs are used in protein-protein interactions across all superfamily members with known structure (see residues highlighted in red in Figure 2.12c). The interactions are with the GTPase phosphate-binding loop (p-

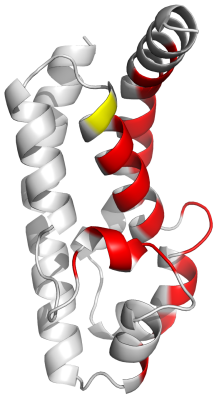
loop), the GTPase switch regions, and the bound nucleotide.



(a) *Pseudomonas aeruginosa* Exos toxin in complex with human Rac.



(b) *Salmonella enterica* subsp. *enterica*, serovar Typhimurium LT2 Tyrosine Phosphatase SPTP in complex with human Rac.



(c) GAP domain superfamily representative with mapped interface residues from different functional families.

Figure 2.12: Example of a small superfamily with limited coverage of protein-protein interfaces, the Bacterial GTP-ase Activating Protein (GAP) domain superfamily (CATH code 1.20.120.260). The GAP domain is displayed in grey cartoon and in the same orientation. Interacting domains are shown in other colours. In Subfigures 2.12a and 2.12b, the protein-protein interface residues on the GAP domain are coloured black. Subfigures 2.12a and 2.12b display PDB entries 1he1 and 1g4u, respectively. Subfigure 2.12c displays the superfamily representative in grey cartoon and inherited protein-protein interaction sites, which are coloured according to the percentage of functional families that have an interface residue at a given position, using the following colour scale: 0 in grey, 1-20% in blue, 20-40% in green, 40-60% in yellow, 60-80% in orange and 80-100% in red).

Example of a functionally diverse superfamily with common functional sites The "Two Dinucleotide Binding Domains" Flavoproteins (tDBDF) superfamily (CATH code 3.50.50.60) is a very large and diverse superfamily with 127 s60 clusters and 22 SSG5 clusters. The superfamily representative however has low functional site coverage as only 37% of residues are mapped to protein-protein interface binding sites, where each mapped position comes from one or more functional families. Figure 2.13 highlights the position of this superfamily relative to others in terms of functional site coverage and sequence diversity.

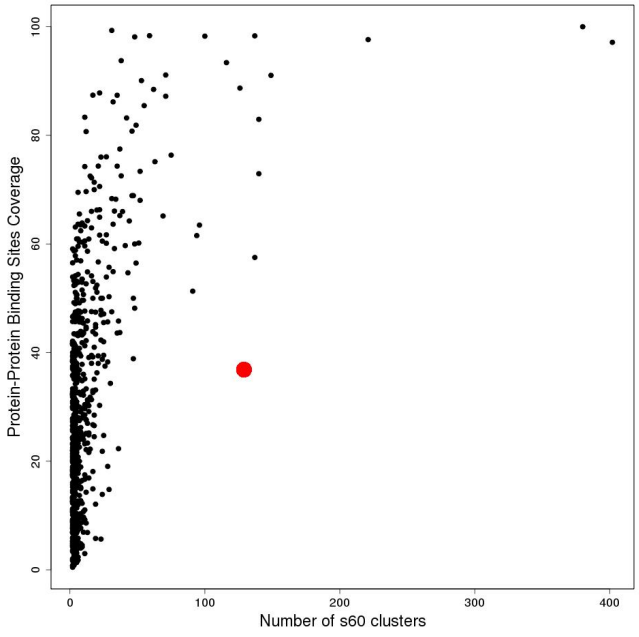


Figure 2.13: The "Two Dinucleotide Binding Domains" Flavoproteins domain superfamily (CATH code 3.50.50.60) is a large superfamily with low functional site coverage and high sequence diversity. The position of the superfamily in the plot is highlighted in red.

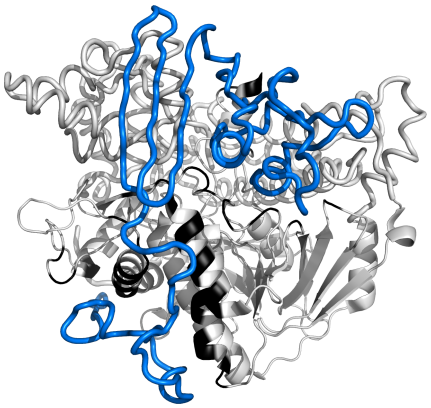
Members of this domain superfamily are known to catalyse multiple types of oxidation/reduction reactions in energy metabolism, apoptosis, maintenance of redox homeostasis and cellular signalling (Ojha *et al.*, 2007). A variety of small molecules and proteins are used as substrates. The superfamily gets this name as relatives belong to proteins using two dinucleotide binding Rossmann-fold domains. The N-terminal domain, termed the FAD domain, typically binds a flavin dinucleotide

(FAD) and the C-terminal domain is known as the NADPH domain, which binds a pyrimidine nucleotide. Both of these domains have a modified nucleotide-binding fold (Bieger and Essen, 2001).

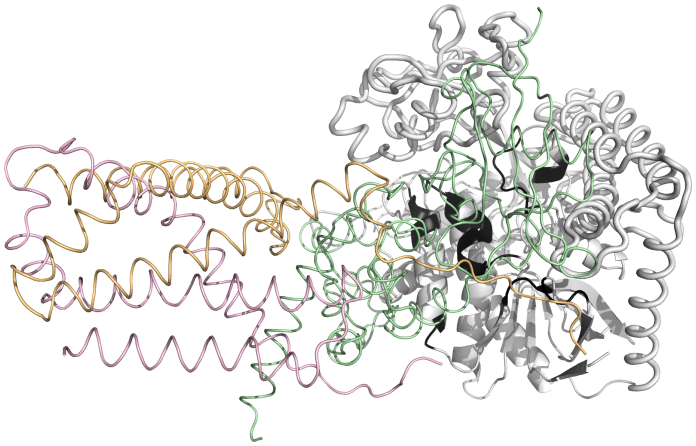
The examples in Figure 2.14 show these two domains in the context of homodimeric enzymes. Members of the tDBDF superfamily contain disulphide oxidoreductase enzymes which include glutathione reductase, thioredoxin reductase, adenylylsulphate reductase (Subfigure 2.14a), and quinol-fumarate reductase (Subfigure 2.14b). Each of these enzymes are made up of two identical monomers, which each have four domains. In some of the figures, the crystallised structures lack some of the domains.

The FAD cofactor is positioned between the FAD and the NADPH domains and the main contacts between the domains are via the isoalloxazine ring system of the FAD cofactor (Bieger and Essen, 2001). The superfamily representative structure is therefore colour-coded grey in this region due to the lack of protein-protein interface residues.

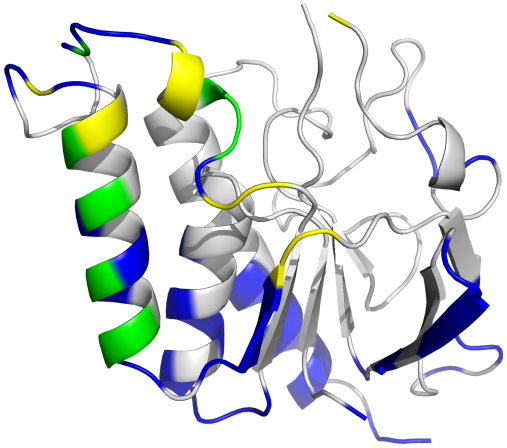
The superfamily representative domain represents the FAD domain and the yellow and green regions shown on its alpha-helices (Figure 2.14c) correspond to interfaces between the monomers in the homodimer. Figure 2.15 illustrates the basic structure of the homodimer using the protein structure from which the superfamily representative was taken. This PDB structure (1FL2) only consists of two domains from a monomer. The majority of interfaces were reported to occur between the two FAD domains, hence the frequently used yellow region in Subfigure 2.14c. These interactions are reported to stabilise the dimers formed by relatives in this superfamily (Bieger and Essen, 2001).



(a) tDBDF domain from the subunit A of Adenylylsulfate reductase from *Archaeoglobus fulgidus* in complex with subunit B.



(b) tDBDF domain from the flavoprotein subunit of Quinol-fumarate reductase, from *Escherichia coli*, in complex with other subunits.



(c) tDBDF superfamily representative with mapped interface residues from different functional families.

Figure 2.14: Example of a large and diverse superfamily with limited coverage of protein-protein interfaces. This is the "Two-Dinucleotide Binding Domains" Flavoprotein (tDBDF) superfamily (CATH code 3.50.50.60). The tDBDF domain is displayed in grey cartoon and the same orientation. The interacting partners are represented as coloured cartoon tubes in Sub-figures 2.14a (PDB entry 1jnr) and 2.14b (PDB entry 1kf6). Interface residues on the tDBDF domain are coloured black. Subfigure 2.14c follows the same colour scheme as described in Figure 2.12.

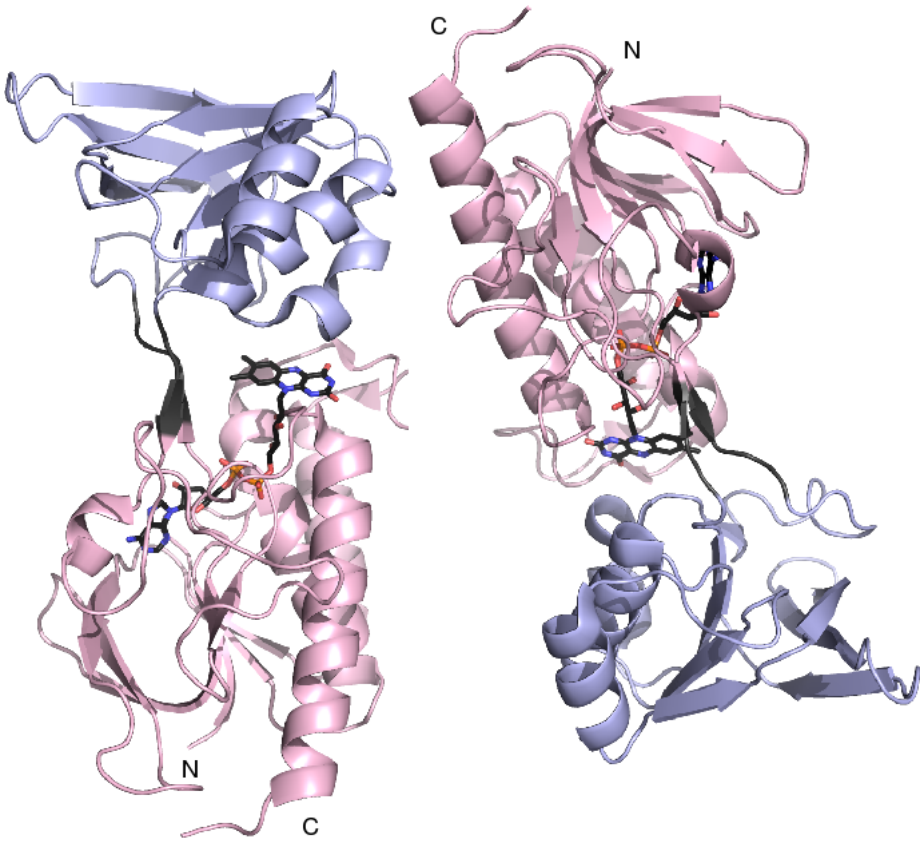


Figure 2.15: The basic structure of the homodimer formed by the two domains in the "Two-Dinucleotide Binding Domains" Flavoprotein (tDBDF) superfamily (CATH code 3.50.50.60) (adapted from Bieger and Essen (2001)). The FAD domain is highlighted in pink and the NADPH domain in blue. The FAD cofactor is highlighted in black.

Example of a functionally diverse superfamily with diverse functional sites The NAD(P)-binding Rossmann superfamily (CATH code 3.40.50.720) is extremely diverse with high functional site coverage, $\sim 97\%$ protein-protein interface coverage, and high sequence diversity, 402 s60 clusters. It also shows high structural diversity with 52 SSG5 clusters. This superfamily is highlighted as a large red dot in the top right-hand corner of Figure 2.16.

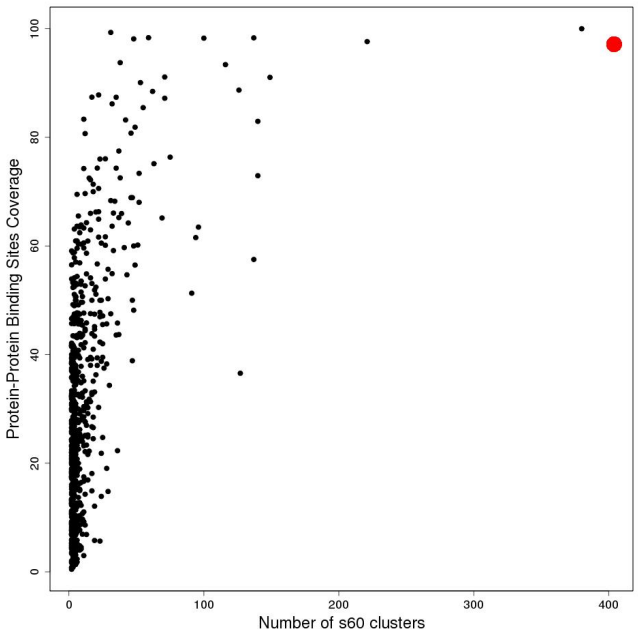
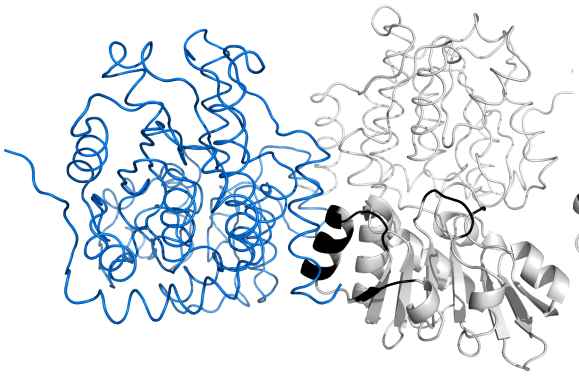
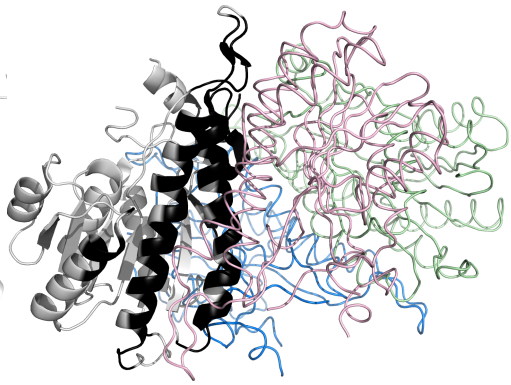


Figure 2.16: The NAD(P)-binding Rossmann domain superfamily (CATH code 3.40.50.720) is highlighted in red to show its position within all superfamilies in terms of functional site coverage and sequence diversity. This superfamily has high functional site coverage and high sequence diversity.

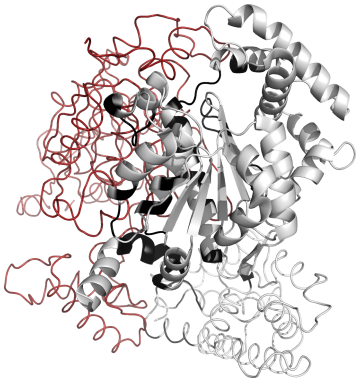
As suggested by their name, these domains bind the coenzyme nicotinamide adenine dinucleotide (NAD) and they also bind a large selection of catalytic domains, assigned to multiple CATH superfamilies (Bashton and Chothia, 2002). It is the bound catalytic domain that determines substrate specificity and the enzymatic reaction. Bashton and Chothia (2002) describe four patterns of connectivity, which are observed in some of our examples. The catalytic domain can be fused to the NAD(P)-binding Rossmann domain at the N- or C-terminus, or in the middle of the Rossmann domain sequence (Figures 2.17c and 2.17e). The NAD(P)-binding Rossmann domain can also be fused in the middle of the catalytic domain sequence (Figure 2.17a). Many of the examples in Bashton and Chothia (2002) show the catalytic domain positioned close to the yellow region in Figure 2.17f. The yellow region corresponds to interfaces with other chains in the protein complex.



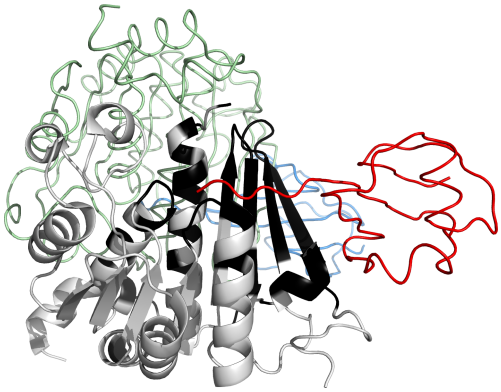
(a) Human S-adenosylhomocysteine hydrolase homodimer.



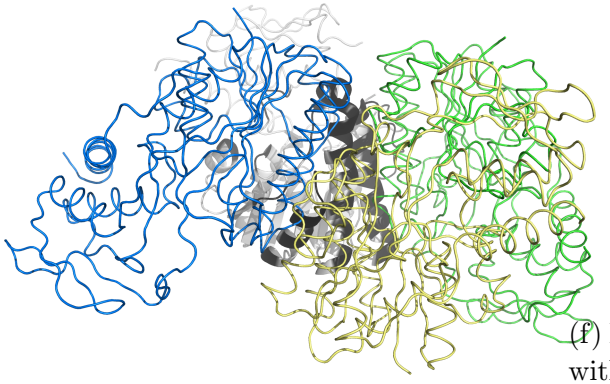
(b) Rat brain 3-hydroxyacyl-CoA dehydrogenase homodimer.



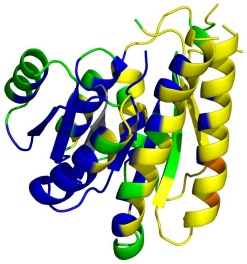
(c) Human NEDD8-activating enzyme E1 regulatory subunit in complex with the catalytic subunit.



(d) Sulfur carrier protein ThiF adenylyltransferase in a heterotetrameric complex with Sulfur carrier protein ThiS, from *Escherichia coli*.



(e) Homotetramer of GDP-D-mannose dehydratase from *Aquifex aeolicus VF5*.

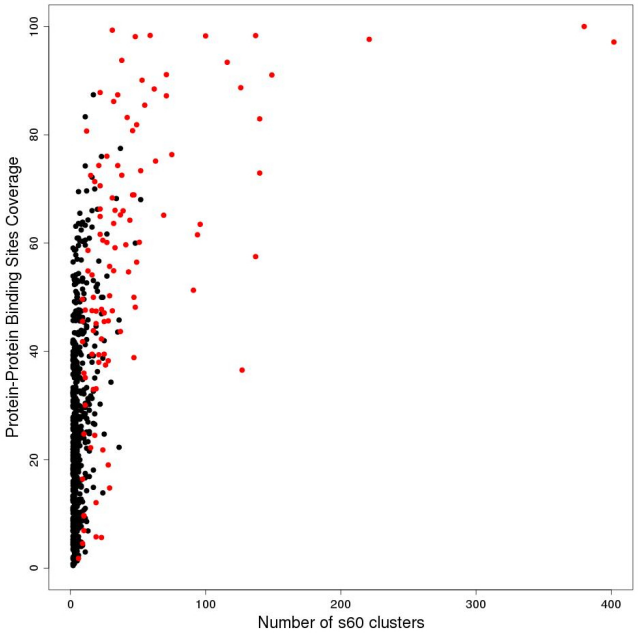


(f) Rossmann superfamily representative with mapped interface residues from different functional families.

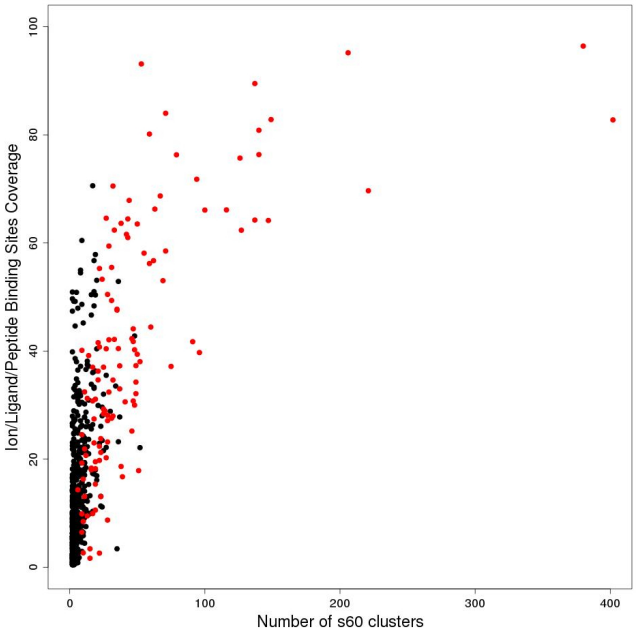
Figure 2.17: Example of a large and diverse superfamily with large coverage of protein-protein interfaces. This is the NAD(P)-binding Rossmann superfamily (CATH code 3.40.50.720). The Rossmann domain is displayed in grey cartoon and the same orientation. Extra domains from the same chain are displayed as grey traces. Interacting partners are displayed as coloured traces. Interface residues on the Rossmann domain are coloured black. Sub-figures 2.17a through to 2.17e display PDB entries 1a7a, 1e3w, 1tt5, 1zud, and 2z1m, respectively. Subfigure 2.17f shows the representative with residues coloured according to the same colour scheme as described in Figure 2.12.

2.3.5 Examining the relationship between structural and functional diversity

A structurally diverse superfamily is defined here as a superfamily with at least five SSGs, generated with a cutoff of 5 Å. Figure 2.18 shows that structurally diverse superfamilies are more likely to have larger functional site coverage values.



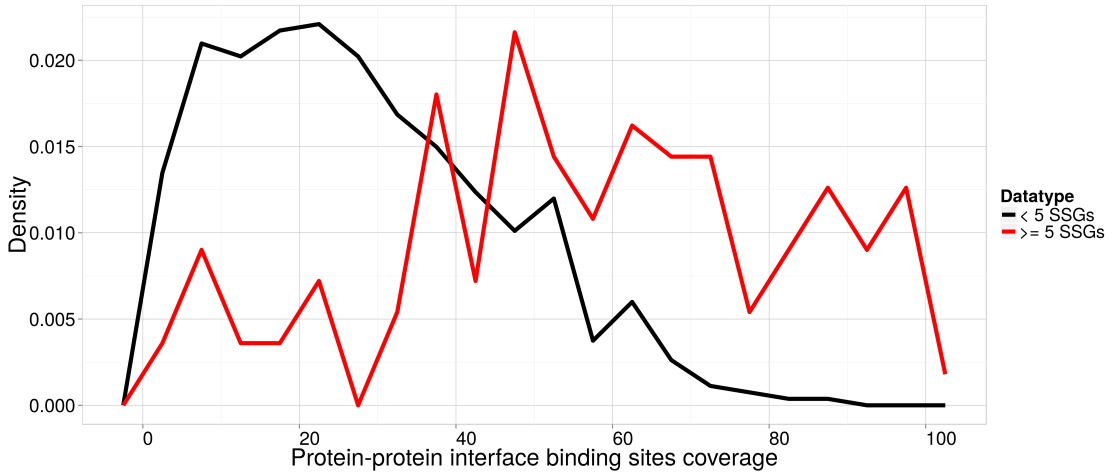
(a) Protein-protein interface coverage



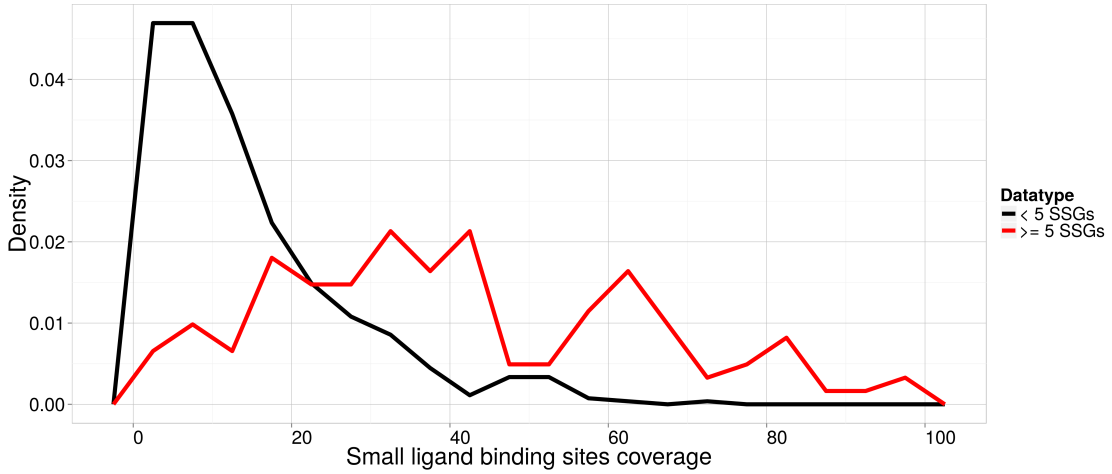
(b) Small ligand binding site coverage

Figure 2.18: Functional site coverage versus superfamily diversity. Structurally diverse superfamilies are shown in red, i.e. those with at least five structural clusters, where a cutoff of 5Å was used to generate the clusters.

The distributions of these functional site coverage values in structurally diverse and structurally coherent superfamilies were found to be significantly different, using an unpaired, one-sided Wilcoxon Rank-Sum test, with a p-value less than 2.2×10^{-16} (see Figure 2.19).



(a) Protein-protein interface binding sites coverage



(b) Small ligand binding site coverage

Figure 2.19: Number of superfamilies versus functional site type coverage. The black line represents superfamilies with less than five structural clusters, and the red line represents superfamilies with at least five structural clusters.

2.3.6 Making functional family and functional site information available on the CATH website

Following the processing and validation of the functional families, the information was published in Sillitoe *et al.* (2013) and Lees *et al.* (2014) and it has also been made available from the CATH website online. A list of the latest functional families from CATH version 4.0 (established in June 2014) is provided for each CATH superfamily, together with information on: the number of sequences in their alignment after filtering the information content of the functional family alignment, the repre-

sentative domain (which is chosen to be a structural domain where available), and whether or not there are any functional sites in the PDB domain (see Figure 2.20). This information content score is defined by the Diversity of Positions (DOPs) score, which is calculated by Scorecons, and reflects the proportion of diverse sequences in the alignment and the number of highly conserved positions. A functional family tree (bottom-left of Figure 2.20) describes the structural relationships between the functional families by illustrating which SSG cluster they are within.

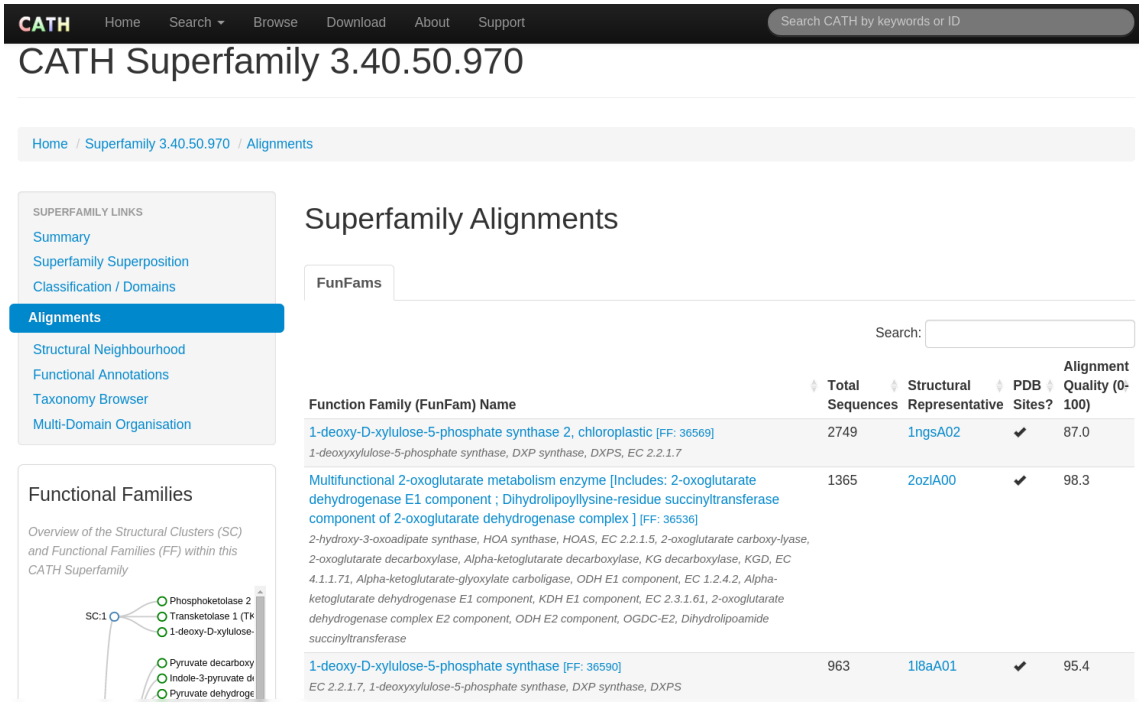


Figure 2.20: CATH webpage screenshot showing a partial list of the functional families in the Thiamine diphosphate-dependent superfamily (ID: 3.40.50.970).

Selecting a functional family name from this list, or from the functional family tree, takes the user to pages summarising functional annotations for the family, together with information on species distributions. If the functional family has a representative with a PDB structure, this is displayed alongside the multiple sequence alignment (see Figure 2.21). In the functional family sequence alignment, highly conserved positions are highlighted in green and displayed on the representative structure.

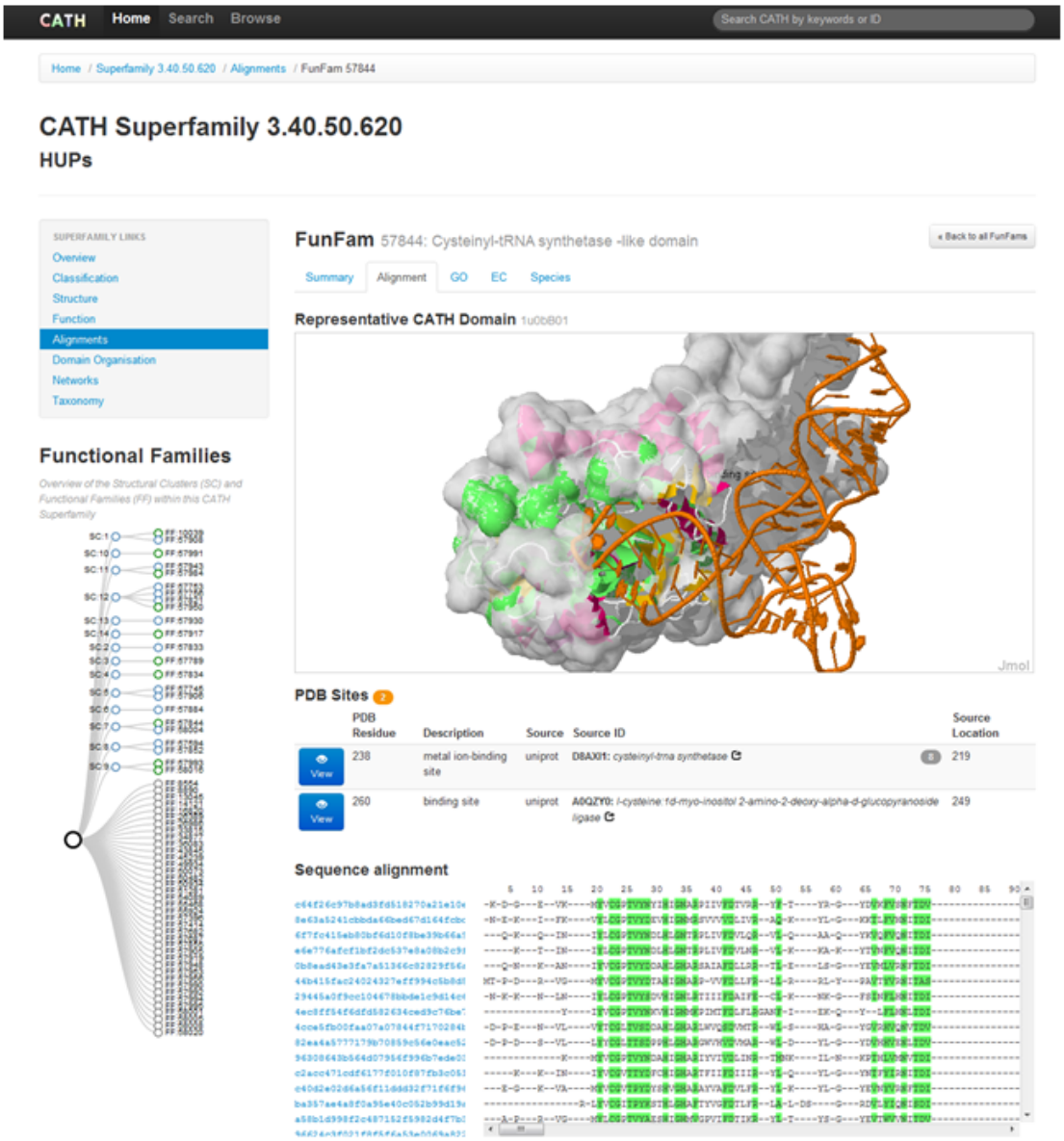


Figure 2.21: CATH webpage screenshot showing the 'Cysteinyl-tRNA synthetase' functional family representative domain information. The domain is classified within the HUPs superfamily (ID: 3.40.50.620). Functional residue sites are listed under 'PDB Sites'.

2.4 Conclusions and future work

This work identified and quantified functional site diversity in CATH domain superfamilies. Different types of functional sites were identified, ranging from catalytic sites to different types of ligand binding sites and protein interface sites.

The removal of sequence fragments from the FunFam$_{GO}$ and FunFam$_{SEQ}$ functional families was shown to increase the family alignment quality. This was validated by an increase in the proportion of conserved CSA and IBIS functional residues. There are approximately 100 FunFam$_{GO}$s and FunFam$_{SEQ}$s with no conserved residues following sequence fragment removal, which suggests that the family alignment quality can still be further improved. For example, these FunFams may contain relatives that use different catalytic residues to catalyse different enzymatic reactions and therefore the families should be further divided. Although the FunFam$_{GO}$s show significant enrichment in catalytic, nucleic acid binding, and small ligand binding functional site residues, in the FunFam$_{SEQ}$s, significant enrichment is seen for all four types of functional site residues. This is likely to be a result of the FunFam$_{SEQ}$s comprising more functionally related sequences, which is also reflected in the structural coherence analysis where FunFam$_{SEQ}$ relatives superpose with lower RMSD values in comparison to FunFam$_{GO}$ relatives. Examples of superfamilies with extreme structural and functional site diversity have been identified and discussed.

Our results show that whilst catalytic sites are generally identified within a small, restricted number of spatial positions, i.e. they tend to be located in a site common to all members of the superfamily, most structurally diverse superfamilies are very flexible in the spatial locations of other functional sites. Protein-protein interfaces display the most flexibility in spatial location. The locations of small ligand binding sites are more varied within a superfamily than catalytic sites, however they are not as diverse as protein-protein interfaces.

This chapter confirmed previous findings of preferred locations for catalytic functional site residues (Dessailly *et al.*, 2010). In enzymes, the active site is usually found in a large and deep surface cleft (Laskowski *et al.*, 1996). Further analysis would be needed to confirm whether the preferred catalytic residue locations in this chapter are found in such a cleft. Using a large cleft can be advantageous as it can: maximise the interactions between the protein and its substrate, the sub-

strate can optimally position itself for catalysis, and the substrate is buried in a solvent-free environment which allows it to form the electrostatic forces required for catalysis (Laskowski *et al.*, 1996).

This chapter also reported that the majority of superfamilies have a small ligand binding site coverage of less than 50%. The term 'small ligand' here refers to small organic compounds, peptides and ions. The relatively low coverage we observe may be a result of small ligands being used as substrates to bind in enzyme active sites, which have shown to be constrained in their functional site location. The other small ligands not being used as substrates will be less constrained in their binding sites and most parts of a protein surface may be able to evolve the required properties to bind these. Relatives in different organisms for example often bind different metal ions and different cofactors and their protein surfaces clearly evolve to bind these different molecules in different ways.

Protein-protein interactions (PPI) allow proteins to form multi-protein complexes that perform a number of essential biological processes including translation, transcription, signal transduction and gene regulation (Thompson *et al.*, 2012). Thompson *et al.* (2012) report that interfaces are larger and flatter than other types of functional sites with an average surface area of 1940 +- 760Å. It would therefore take relatively few relatives with different PPI interfaces to result in a high coverage of data on the superfamily representative. These larger sites could help suggest why we are observing such large superfamily coverage.

This work also supported the previous study of Dessailly *et al.* (2010) in showing that structurally diverse superfamilies tend to have a high coverage of interfaces and suggested that diverse functional relatives have different domain or protein partners. A hypothesis which is also supported by studies showing that relatives in different species do not necessarily have the same interaction partners (Reid *et al.*, 2010). Interestingly however, there is some overlap between different family interfaces in a majority of superfamilies. Further analysis would be needed to ascertain whether this overlap is larger than expected by chance. To determine whether this is due to

some structural characteristic of the surface or a constraint imposed by the location of the cofactor would require further study.

Protein interactions also exhibit a wide range of affinity values, ranging from strong picomolar dissociation values in stable complexes through to weak millimolar dissociation values in transient complexes. As it is difficult to study and crystallise transient structures, these are under-represented in the PDB (Thompson *et al.*, 2012) and therefore in this work.

Due to structural diversity within the functional families and superfamilies studied there may be large insertions, or embellishments, that are not shared across the members. Functional site residues identified within such embellishments therefore may not be mapped onto a representative that does not also have an insertion at the same place, resulting in potentially missing functional site information. As the functional family and superfamily representative was chosen as the most structurally similar domain, this method should prevent the chosen representative as being one with unique embellishments which would minimise the effect of missing any functional site mappings and therefore under-estimating functional site diversity. Further work would need to be done to examine the extent of any missing functional site information.

Chapter 3

Understanding the Mechanisms of Functional Diversity in Enzyme Domain Superfamilies

3.1 Introduction

3.1.1 Background

Enzymes are powerful, specific, and essential proteins that catalyse chemical transformations. They are used by cells to catalyse the production of molecules required during growth, repair, maintenance and death. Nearly all enzymes are proteins, with the exception of some catalytic RNA molecules, and they carry out catalysis within an active site. There are thousands of enzymes that together provide a wide diversity of chemical reactions due to their ability to specifically bind and act on a wide range of substrates (Berg *et al.*, 2002).

Emil Fisher in 1890 first suggested that an enzyme and its substrate fit together like a lock and a key, i.e. the two are highly specific for each other and have a one-to-one relationship. This model was updated in 1958 by Daniel Koshland, who hypothesised that an enzyme is structurally flexible upon binding the substrate and produced the induced-fit model of enzyme-substrate binding (Koshland, 1958). This model reflects the fact that the substrate does not have to be an exact fit to the enzyme's active site, which is the reason many enzymes can bind more than one substrate and potentially catalyse a variety of chemical reactions.

Chemical reactions can be greatly accelerated by the presence of an appropriate enzyme. Carbonic anhydrase catalysing the hydration of carbon dioxide is one of the fastest known enzymes. This enzyme can hydrate carbon dioxide at rates up to 1 million times a second, which is over 6 million times faster than without an

enzyme (Berg *et al.*, 2002).

As well as a substrate, a number of enzymes also require a cofactor to perform their activity. Cofactors are small molecules that can be divided between metal or non-organic molecules, and small organic molecule subgroups. They are often derived from vitamins and bind either tightly or loosely to the enzyme (Berg *et al.*, 2002). Such data is stored in resources such as the CoFactor database at the EMBL-EBI, which contains manually-curated data taken from the literature on 27 small organic cofactors, as well as automatically-derived information (Fischer *et al.*, 2010).

3.1.2 Characterising the enzyme active site

The active site of an enzyme, typically found in a large surface pocket, plays a key part in catalysis. It allows the ligand to bind in a solvent-free environment, which stabilises polar transition states through neighbouring charged residues or metal ions. As the active site is vital for the performance of an enzyme, the structural location and the amino acid type of the catalytic residues have been shown by a number of studies to be generally conserved between enzyme relatives performing the same chemical reaction. A well-known example is the completely conserved catalytic triad, Asp-His-Ser found in trypsin, chymotrypsin, and elastase, which are homologues belonging to the serine protease superfamily. All three hydrolyse a peptide bond, however they bind and act on different substrates (Krem *et al.*, 2000), and produce different products. More recent studies have shown that it is not uncommon for catalytic residues to change in location within the active site or to change in their amino acid type between relatives in a functionally diverse homologous superfamily (Dellus-Gur *et al.*, 2013).

During the last decade, the amount of active site and catalytic residue data available in the literature has risen. This data has been collected and published online through resources such as the Catalytic Site Atlas (CSA). The CSA stores information on catalytic site residues derived from the literature, and also from ho-

mology searches, and defines them as either, a residue: 1) with direct involvement in the catalytic mechanism, 2) effecting another residue or water molecule directly involved in the catalytic mechanism, 3) that stabilises a transition-state intermediate, or 4) that exerts an effect on a substrate/cofactor aiding catalysis. Residues involved in ligand binding are excluded unless they are involved in one of the four definitions (Bartlett *et al.*, 2002; Porter *et al.*, 2004).

Bartlett *et al.* (2002) analysed 615 catalytic residues in 178 enzyme active sites and reported an average of 3.5 catalytic residues per enzyme. They described most catalytic residues (65%) as being charged (His, Arg, Lys, Glu, Asp), 27% as being polar (Gln, Thr, Ser, Asn, Cys, Tyr), and 8% as being hydrophobic (Gly, Ala, Val, Leu, Ile, Phe, Trp, Pro, Met). Charged residues were abundant as they are required during catalysis to create electrostatic forces, which enable the movement of protons and electrons between the acceptor and donor, and they also provide charge stabilisation. Polar residues also play a part in forming this environment. Bartlett *et al.* (2002) concluded that there was no correlation between the abundance of a given residue type and its contribution to catalysis. They also showed that 50% of the analysed catalytic residues occurred in loop regions, which is more than expected by chance, and described lower distributions for alpha helices (28%) and beta-strands (22%).

3.1.3 Divergence of function across protein domain superfamilies

Neidhart *et al.* (1990) was the first study showing that related enzymes did not necessarily catalyse identical reactions. They solved the structures of mandelate racemase (MR) and muconate lactonising enzyme (MLE), through X-ray crystallography, to find that despite being homologous enzymes and therefore sharing a common (TIM-barrel) structural fold, they catalysed completely different chemical reactions.

Many other pairs of homologous enzymes have since been shown to catalyse

different chemical reactions using the same or different catalytic machineries. Structural determination of protein structures through X-ray crystallography has been particularly important in identifying and understanding how such events occur. The Triose Phosphate Isomerase (TIM) barrel fold is an example of a structure supporting many diverse functions. TIM is the central enzyme in the glycolytic pathway and it was the first enzyme found to contain the structural motif of eight parallel beta-strands, each one followed by an alpha-helix. The eight strands of the beta-sheet form a barrel-like centre and the helices cover the barrel-like structure, and this scaffold has frequently been used to perform different enzyme chemistries (Petsko *et al.*, 1993).

3.1.3.1 Divergence of chemistry and substrate specificities in protein superfamilies

Studies have been carried out to explore the nature of functional divergence within multiple superfamilies. Babbitt and Gerlt (1997) discuss a small-scale study of four enzyme superfamilies whose members share the same TIM barrel structural fold, or scaffold, but that catalyse different chemical reactions: the Enolase, N-Acetylneuraminate Lyase, Crotonase, and the Vicinal Oxygen Chelate (VOC) Fold superfamilies. Babbitt and Gerlt (1997) hypothesised that there is a common chemical reaction step to all superfamily members. They discovered a higher number of catalytic functions that could be supported by a single structural scaffold than previously determined. They also found evidence to support that adding new catalytic residues to an active site, whilst conserving the catalytic machinery needed to catalyse the common reaction step, leads to the divergence of enzyme function. It was suggested that this common reaction step was involved in stabilising chemically-similar transition states and intermediates (Babbitt and Gerlt, 1997).

Due to the increase in information associated with these superfamilies, for example enzyme reaction mechanism data from wild type and mutant studies, sequence data from genome sequence projects, and structural data, Gerlt and Babbitt (1998)

went on to characterise the same four superfamilies further. They also wanted to better understand whether they could assign an enzyme with unknown function to a superfamily using only reaction mechanism information. They discovered that defining a common reaction step between members of these four superfamilies was more complicated than previously described. The shared step was not always involved in the stabilisation of transition states or intermediates but could instead be involved in another part of the reaction, as shown in members of the Vicinal Oxygen Chelate (VOC) fold superfamily. Members of this superfamily did not catalyse an obvious shared reaction step but instead used a divalent metal ion to stabilise a negative charge that occurred in a number of different reaction steps. This example required the use of structural and genetic information as well as reaction mechanism information to deduce the common reaction step between superfamily members (Gerlt and Babbitt, 1998). These results suggested that for large-scale studies and for very functionally diverse superfamilies, in order to assign an unknown enzyme to a superfamily one needs to use a number of different methods rather than simply looking at reaction mechanism steps.

Todd *et al.* (2001) analysed 28 enzyme superfamilies known to bind substrates and found a conservation of chemistry but surprising variations in the diversity of substrates that were acted on by members of the same superfamily. Different substrates were acted on by different superfamily members in all but one superfamily. In a high proportion of these superfamilies (20 out of the 28) the substrate specificity was completely diverse. They varied in their size, chemical properties, and/or in their structure scaffolds (e.g. if they were aromatic versus linear-chain hydrocarbons). Enzymes in 6 of the 28 superfamilies bound a common substrate type, for example nucleic acid, sugars, or phosphorylated proteins. The two remaining superfamilies had little or no variation in their substrate binding. The phosphoenolpyruvate binding domains superfamily consist of members that only bind the substrate phosphoenolpyruvate, and the ribulose-phosphate-binding superfamily whose substrates, and products, had to have a glycerol or ribulose-phosphate group. This

is because three out of the five superfamily members catalyse sequential steps in the biosynthesis of tryptophan and the product of one reaction is the substrate of the next (Todd *et al.*, 2001).

More recent analyses of functional divergence within highly divergent homologous superfamilies, but on a much larger scale, have shown that some large superfamilies contain diverse relatives that catalyse completely different chemical reactions (Dessailly *et al.*, 2010). Functional divergence within superfamilies can be explored through the FunTree resource (Furnham *et al.*, 2012). FunTree reconstructs phylogenetic trees for selected CATH superfamilies and highlights functional diversity through publicly available data on enzyme reaction mechanisms from MACiE and catalytic residue data from the CSA. Furnham *et al.* (2012) calculated the distribution of the number of EC terms for each of FunTree's 276 CATH superfamilies, and found that while 49 superfamilies (17.75%) only had one associated EC number, there are some superfamilies that have numerous, for example the NAD(P)-binding Rossmann-like domain was highly functionally diverse with 223 unique EC numbers. When they ordered the superfamilies by the number of associated sequences, it was discovered that 10% of enzyme superfamilies accounted for 849 unique functions (i.e. EC numbers), with an average of 35 EC numbers per superfamily. The remaining superfamilies had an average of 6 EC numbers. The fourth hierarchical level of an EC number was used to gauge promiscuity; 177 superfamilies showed functional diversity at this fourth level and in 150 superfamilies at least half of their EC number diversity was due to changes at the fourth level of EC. A change at the third level of EC is defined here as a proxy for a change in enzyme chemistry; 176 superfamilies ($\sim 64\%$) have at least one member with an EC that is diverse at the third hierarchical level (Furnham *et al.*, 2012).

3.1.3.2 Mechanisms of functional divergence

Large scale analyses on the mechanisms of functional divergence were carried out by the Thornton group on 31 enzyme superfamilies classified in the CATH database (Todd

et al., 2001). They reported that new enzyme functions are frequently due to gene duplication events and incremental mutations. These incremental mutations may lead to differences in the catalytic machinery of an active site. Other mechanisms of functional divergence discussed included: oligomerisation, where two or more copies of the same protein, or at least one copy of two or more different proteins, form a protein complex; gene fusion; gene recruitment, where the function of the gene product depends upon the environment in which it is expressed; gene fission; alternative gene splicing; exon shuffling through intronic recombination; post-translational modifications; and the presence of different metal ions and/or cofactors in the active site (Todd *et al.*, 1999; Buljan and Bateman, 2009).

Indels and structural embellishments Reeves *et al.* (2006) described another method of functional divergence through the insertion or deletion of amino acid residues (indels). They manually inspected 48 structurally diverse domain superfamilies in CATH and found short secondary structural insertions, or embellishments, typically less than 15 residues long, that were usually discontiguous at the sequence level but often co-located in 3D. These insertions often resulted in a modified active site structure or a modified protein surface that promoted more diverse domain and protein interactions. As the majority of the indels were found in the loop regions in between secondary structure elements, these mutations were shown to be tolerated because they did not affect the stability of the fold. Residue mutations that do fall in the structural core and involve changes in the size of the amino acid residues are generally compensated by shifts in the secondary structure to maintain tight residue packing (Reeves *et al.*, 2006).

Later studies also showed that indels causing large length differences within domain superfamilies could affect the function of the protein and promote functional diversity (Sandhya *et al.*, 2009). Out of the 353 domain superfamilies analysed, those with at least 4 members had 20% of their domains exhibiting large length variation due to indels, i.e. a domain was over 30% longer than the average domain size. Sandhya *et al.* (2009) reported that these longer domains can: confer extra

thermal stability, influence subunit interactions and therefore affect the quaternary structure, affect substrate specificity, and generate new interaction interfaces. A total of 64 (18.13%) superfamilies had at least 75% of their members showing large length variation and were termed 'length-deviant' superfamilies. Within these 64 superfamilies, domain structural repeats were found to be common in 27 of them. Diverse domain contexts were found where associations were observed with multiple copies of the same or different domain superfamilies, and functional interactions were found with large numbers of different protein domains.

3.1.3.3 Balance between maintaining protein stability and mutations which drive functional change

Dellus-Gur *et al.* (2013) hypothesised that there are two main factors in the evolution of protein function: genetic robustness, where mutations are accumulated but do not affect function; and innovability, the acquisition of new functions that have diverged from the original function. Protein stability provides tolerance to amino acid mutations, which promotes robustness. For example a highly-ordered, well-packed protein has a high stability threshold, which allows more destabilising mutations to accumulate (than if the protein structure had a low stability threshold), and this in turn promotes innovability. However an increase in stability also means a decrease in conformational plasticity, which is often required by new functionalities.

Dellus-Gur *et al.* (2013) looked for evidence of a trade-off between the levels of innovability and stability to better understand the evolution of function. They analysed single-domain proteins found in the LUCA (Last Universal Common Ancestor) with a high-resolution crystal structure (i.e. < 2.5Å). 43 PDB structures, with a total of 28 different EC numbers, from only eight folds in CATH (i.e. the same class, Architecture, and Topology classification) had enough representative examples. The structures belonging to each fold group were structurally aligned using MUSTANG (Konagurthu *et al.*, 2006) and scaffold positions were defined as $\geqslant 70\%$ conservation of secondary structure in the alignment. The scaffold represents the

secondary structure elements shared by all enzymes with a given fold.

They found that in protein folds that have carried out a single function throughout evolution, most of the active site residues (e.g. 60% in dihydrofolate reductase) are found within the highly stable, and therefore rigid, structural scaffold. However, multi-functional protein folds such as the TIM barrel were found to have only around 20% of their active site residues in the structural scaffold and the rest were found in loops that are able to be more plastic. This is an example of a trade-off between innovability and structural stability as it allows the structural scaffold to remain robust, whilst the active site residues evolve at a faster rate to promote new functionality (Dellus-Gur *et al.*, 2013).

3.1.3.4 Multifunctional and promiscuous enzymes

While it was originally thought that one enzyme only binds one substrate, i.e. the lock and key hypothesis by Fischer (1890), it is now understood, as discussed already above, that many enzymes have the ability to interact with more than one substrate. This multiple substrate specificity enables the catalysis of potentially multiple chemical reactions, as each different substrate that is bound will catalyse a single, different chemical reaction. This ability of an enzyme to interact with more than one substrate, can be due to either: enzyme multi-specificity, where an enzyme has broad specificity and can therefore bind multiple substrates in its active site; or promiscuity, where an enzyme coincidentally binds a non-native substrate and catalyses a reaction which it did not evolve to carry out.

The first notion of enzyme promiscuity came in 1976 when Jensen hypothesised that while modern enzymes tend to specialise in one substrate and one reaction, ancient enzymes had very broad specificities (Jensen, 1976). Khersonsky and Tawfik (2010) describe enzyme promiscuity as the additional activities an enzyme has evolved to carry out that are not part of the organism's physiology. By contrast, enzymes such as glutathione S-transferases (GSTs) and cytochrome P450s, which evolved to transform a wide range of substrates are multi-specific, or broad-spectrum

enzymes, and are not promiscuous. Multi-specificity typically involves the same enzyme chemistry being carried out with a range of substrates and therefore only changes at the fourth EC hierarchical, or EC4, level are observed. Promiscuity however typically leads to changes at the first, second and third EC hierarchical levels as very different enzyme chemistries can be performed (Khersonsky and Tawfik, 2010).

Promiscuous enzymes have been shown to catalyse different chemical reactions with the same active site due to a number of influencing factors: 1) the native and the promiscuous functions are controlled by different active site configurations; 2) interactions with different substrates; 3) catalytic residues can act in different protonation states in the native and promiscuous functions; 4) different sub-sites with the same active site can be used; 5) alternative cofactors may be used; 6) water molecules may assist promiscuity as they can buffer charges between the active site and substrate residues, as well as acting as an acid, base or a nucleophile (Khersonsky and Tawfik, 2010).

Many enzymes are known to perform additional functions that are more often related to structural or regulatory functions rather than catalysis, which is referred to as moonlighting (Copley, 2003; Jeffery, 1999). Promiscuity is a form of moonlighting, however it only refers to enzymes that can catalyse different types of chemical reactions. Distinct biological activities are usually seen in proteins that moonlight, which may be a product of the protein's active site or the evolution of other functional sites on the protein. The biological context is frequently a factor in a protein's ability to moonlight, and the function of a moonlighting protein can change due to: a change in the protein's cellular location; a secreted protein losing its enzymatic activity and serving as a growth factor; the cell type in which the protein is expressed; the oligomeric state, e.g. a multimer could have different functional activity to a monomer; the cellular concentration of a ligand, substrate cofactor or product; some proteins having different binding sites for different substrates (Jeffery, 1999).

The E. coli PutA protein is an example of a change in cellular location causing a change in function. PutA has proline dehydrogenase and pyrroline-4-carboxylate

dehydrogenase activity inside the plasma membrane, however in the cell cytoplasm it does not have enzymatic activity and binds DNA as a transcriptional repressor (de Spicer and Maloy, 1993; Muro-Pastor *et al.*, 1997).

3.1.4 Convergence of function across protein domain superfamilies

Aside from functional diversity, functional convergence can occur. Functional convergence has most commonly been studied between superfamilies, i.e. in sets of non-homologous proteins, or protein domains, that carry out the same, or a related, biological function using the same catalytic machineries in different structural scaffolds. For example, early studies by Drenth *et al.* (1972) and Kraut (1977), found that the Ser-His-Asp catalytic triad carried out the same reaction mechanism in two structurally unrelated proteins: chymotrypsin and subtilisin. Zhang *et al.* (1994) later discovered functional convergence in the active sites of tyrosine phosphatases.

Gherardini *et al.* (2007) performed a large-scale analysis to quantify the amount of functional convergence occurring between non-homologous protein domains in the Structural Classification of Proteins (SCOP) database. They looked at two different types of functional convergence, both at the protein domain fold level. They first studied mechanistic analogues, which were enzymes that used the same mechanism, or catalytic machinery, to carry out related reactions, classified using the first three hierarchical levels of the EC number (EC3). They then studied transformational analogues, which were enzymes catalysing the same chemical reaction (identical EC numbers) using different catalytic machineries. Catalytic residues were defined using data from the CSA. They found that the catalytic triad is the most widespread active site, with variation occurring in 23 superfamilies. 15% of the 169 EC3 groups were mechanistic analogues, (i.e. enzymes using the same catalytic machinery to catalyse related chemical reactions) and 4.7% of the 951 EC4 numbers were transformational analogues (i.e. enzymes using different catalytic machineries to catalyse the same chemical reaction). This showed that functional convergence between superfamilies

is not particularly rare and that more often the same catalytic machinery is used.

Since many studies assumed that proteins with the same EC3 classification performed similar reactions, Almonacid *et al.* (2010), studied 95 pairs of enzymes in different superfamilies that catalysed similar chemical reactions to find out whether the EC3 level of annotation is indeed indicative of overall reaction similarity when comparing pairs of functionally analogous enzymes (i.e. non-homologous enzymes carrying out the same enzymatic function). The enzymes pairs were extracted using MACiE entries that had the same EC3 number, and the CATH database was used to define non-homologous enzymes. Enzymes were taken from all four CATH structural classes to provide higher coverage of the data. The overall reaction similarity and the mechanistic similarity (i.e. comparing each step in the reaction mechanism) for each enzyme pair was calculated, and compared to a background data set. Reaction similarity was calculated by comparing the sets of bond changes in the transformations of substrates and products (O'Boyle *et al.*, 2007). They found widespread convergence of reaction mechanism steps, for example, cases of unrelated enzymes that carry out similar, or the same, overall reaction mechanisms (i.e. mechanistic analogues) with similar or different catalytic machineries. However, in contrast to the Gherardini study, they found that different catalytic machineries were found to be used more often. Almonacid *et al.* (2010) concluded that to perform a similar chemical reaction mechanism, an enzyme does not need to use the same catalytic machinery or use a similar 3D environment for its active site.

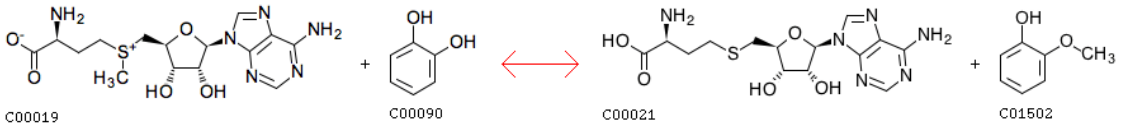
By contrast, functional convergence after evolutionary divergence can also occur within homologous superfamilies, which is when relatives within the superfamily use different catalytic machineries to catalyse similar chemical reactions (Todd *et al.*, 2002). Here, the gene of the common ancestor has been duplicated and it has diverged to produce two primordial enzyme genes carrying out two different functions. These two enzymes then evolve through independent pathways and functionally converge to carry out the same function using different active-site machineries (Todd *et al.*, 2002).

Kuriyan *et al.* (1991) describe an example of functional convergence between homologues in the pyridine nucleotide disulphide oxidoreductase superfamily. They studied the structures of E. coli thioredoxin reductase and glutathione reductase and discovered that they both catalysed the reduction of disulphide bonds, despite using different active site residues. The Zinc-peptide superfamily is another example of functional convergence after evolutionary divergence; the two distantly-related families, Zn-dependent carboxypeptidases (ZnCP) and Zn-dependent aminopeptidases (ZnAP), have a conserved structural core and both families contain three common types of enzyme: 1) proteases, 2) enzymes performing N-deacylation, and 3) enzymes catalysing the N-desuccinylation of amino acids. The ZnCP and ZnAP families have therefore evolved their own reaction specificities independently of each other following their divergence from their common ancestor (Makarova and Grishin, 1999).

3.1.5 Measuring similarities between enzyme chemical reactions

The similarity between two enzyme chemical reactions is typically compared using EC terms and observing whether the numbers are the same at each hierarchical level. The problem with this is that the similarity cannot be quantified using EC numbers and the numbers at each hierarchical level do not necessarily represent reaction similarity if they are close in value. The EC-BLAST algorithm overcomes this issue by automatically calculating the similarity between two enzyme chemical reactions through their associated IUBMB reactions (Rahman *et al.*, 2014). The reaction mechanism similarity is calculated in three different ways, using either bond change, reaction centre, or substructure information. Bond change represents the bonds formed or cleaved, the order changes, and stereo changes. The reaction centre incorporates the atoms directly in contact with the bond being acted on. The small molecular substructures are the small molecule moieties of the reaction (Rahman *et al.*, 2014). Figure 3.1a highlights these three different types of reaction mechanism

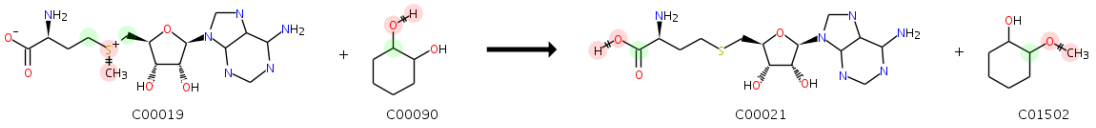
characteristics, based on the chemical reaction of catechol O-methyltransferase (EC 2.1.1.6).



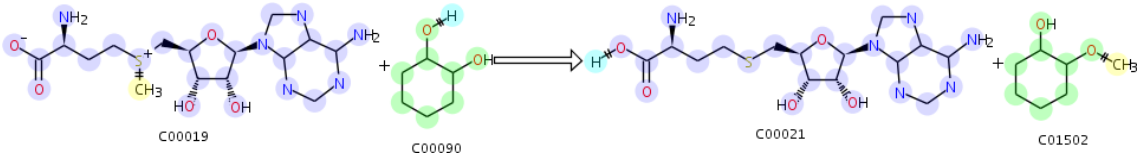
(a) Catechol O-methyltransferase (EC 2.1.1.6) chemical reaction.

Formed/Cleaved (⇌)		Bond Order Changed (=)	Stereo Changes (◁)
H-O	2		
C-O	1		
C-S	1		

(b) The three different types of bond changes that occur and their symbols.



(c) Atoms forming the reaction centres are highlighted with red circles. Interacting atoms are marked with green circles.



(d) Atoms forming common substructures are highlighted with the same coloured circles.

Figure 3.1: Example showing the three different types of reaction mechanism characteristic that are compared to quantify chemical reaction mechanism similarity. Figures have been taken from the KEGG REACTION database (a) and the EC-BLAST web site (b-d).

The MACiE (Mechanism, Annotation, and Classification in Enzymes) (Holliday *et al.*, 2011) resource provides information on enzyme reaction mechanisms. It also provides tool for comparing these reaction mechanisms. The latest version (3.0) contains 335 annotated enzyme reaction mechanisms, which correspond to 321 EC numbers and 372 different CATH superfamilies. Each entry comprises the number of steps in the reaction, the catalytic residues, and the use of any cofactors. The reaction similarity between entries can be compared pairwise or through a matrix. Users can easily search for MACiE entries that have the same EC4 number, and for

CATH domains that share at least one catalytic domain.

Three different characteristics are scored and incorporated to calculate a single overall reaction similarity score for the compared entries: 1) the bond change information, which used the same information as described above, 2) the similarity of the reaction mechanism at each reaction step, where the bond changes across all reaction steps are compared, and 3) the catalytic machinery similarity, where first the complementing catalytic residues are compared, and second their 3D coordinates are compared through the superposition of structures. The catalytic machinery similarity score is a weighted (9:1) combination of these two comparisons. The scores for these three similarities are weighted and summed to give a final overall similarity score (Holliday *et al.*, 2011).

3.1.6 Aims and objectives

This work examines functional subfamilies from over 100 enzyme superfamilies in the CATH database and investigates how enzyme chemistries and catalytic machineries change between the subfamilies across each superfamily. We also examine whether changes in catalytic machineries correlate with changes in reaction chemistries. The location of catalytic residues is also investigated to explore whether there is a preference for a catalytic residue to be within a secondary structure element or a non-structured region, i.e. a loop region.

3.2 Methods

Changes in catalytic machineries have been examined between relatives in enzyme domain superfamilies with regards to their physicochemical properties and structural location. A subset of CATH enzyme superfamilies was created containing all enzyme superfamilies with CSA data for two or more FunFam$_{SEQ}$ functional families. This allowed for comparisons with at least two sets of experimentally-validated catalytic sites.

3.2.1 Identification and mapping of catalytic residues

Using the FunFam$_{SEQ}$ functional families (see definition in Section 2.2.1, page 59), we first identified the superfamily members to be compared. A representative domain with known 3D structure was chosen for each functional family in each superfamily. This representative was calculated as the domain with the highest structural similarity to all other domains within a functional family (see Methods in Section 2.2.5, page 65).

Catalytic residues were identified for each structural domain, within a functional family, with experimentally-identified catalytic residues in the Catalytic Site Atlas (CSA) (Porter *et al.*, 2004) (see Methods in Section 2.2.5, page 64). Catalytic residues were mapped onto the functional family representative as described in Methods Section 2.2.5 (page 65).

Choosing one representative for each functional family reduced the number of domains whose catalytic residues had to be compared, without losing any information as all catalytic residue information for each family was mapped onto the representative. This was thought to be a reasonable protocol as the FunFam$_{SEQ}$s had been shown to be functionally pure in the majority of cases.

3.2.2 Comparing catalytic machinery similarity between functional families

To explore whether catalytic machineries change between functional families in a superfamily, we first mapped catalytic residues between the functional families. This was done in two ways: 1) 'the pairwise alignment-based protocol' by obtaining a structure-based pairwise sequence alignment of two functional family representative domains using SSAP (Taylor and Orengo, 1989) and then selecting known catalytic residues from equivalent positions in the alignment; 2) 'the 3D superposition-based protocol' by superposing pairs of functional family representative structures and mapping catalytic residues that co-locate in 3D to within 5Å.

Subsequently, we compared the physicochemical properties of residues found to be equivalent catalytic residues between the functional family representatives using the physicochemical similarity matrix from McLachlan (1972).

Steps in the protocol are described in more detail in Figure 3.2.

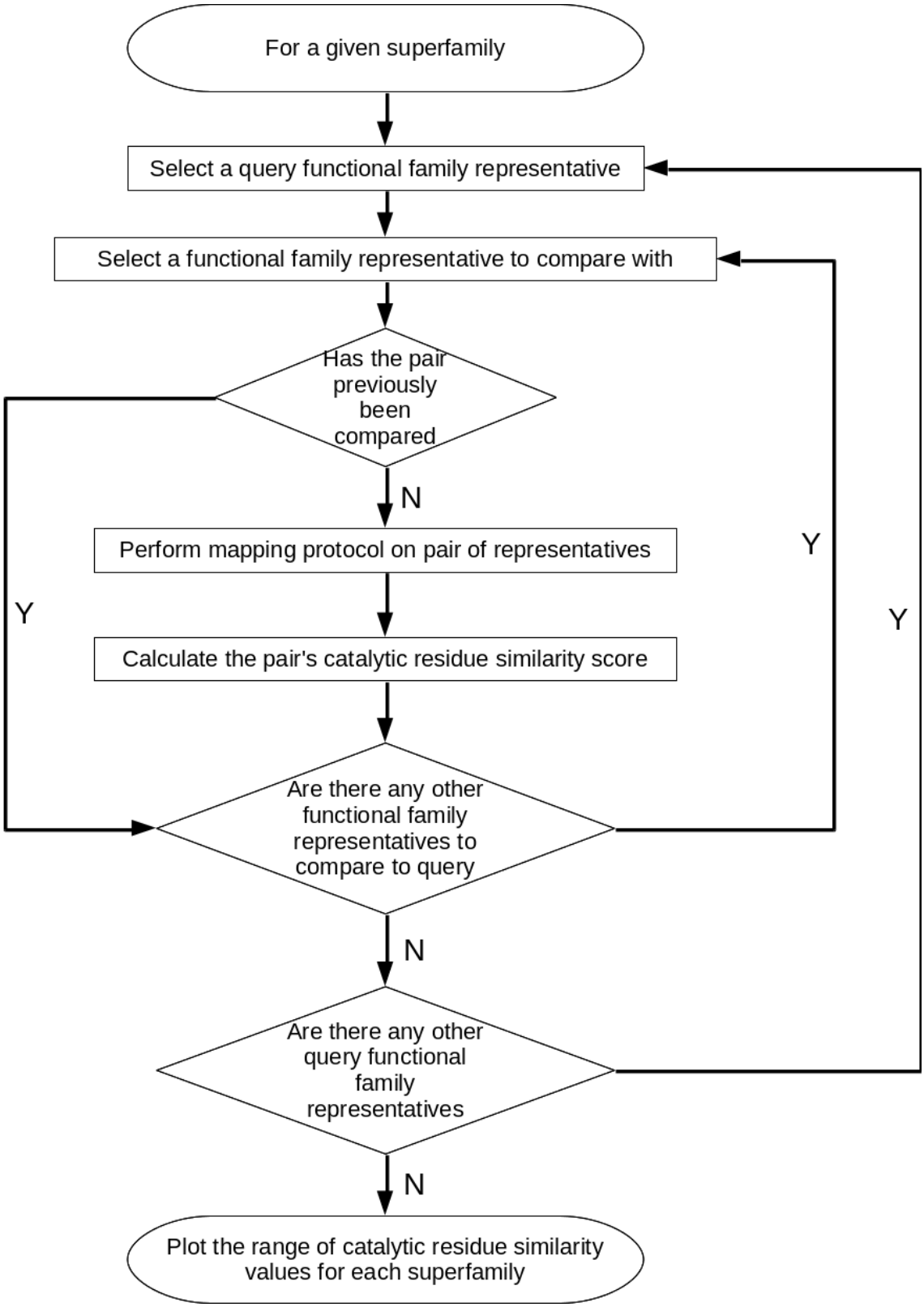


Figure 3.2: The general steps used to score catalytic residue similarity between functional family representative domains across a given superfamily.

3.2.2.1 Comparing catalytic machineries using the pairwise alignment method

For each pair of functional families in a superfamily, a pairwise structure-based sequence alignment was obtained using the structure comparison method SSAP. The pairwise alignment was used to compare aligned residues where at least one of the aligned residues was a known catalytic residue. Functional families were only compared if the representatives aligned well (i.e. with an RMSD less than 5Å).

For each pair of well-aligned functional family representatives, the similarity in aligned catalytic residues was measured using the physicochemical similarity matrix (McLachlan, 1972). This scoring matrix provided a measure of amino acid similarity based upon amino acid polarity, size, shape and charge. Depending on these characteristics, a pair of amino acids was given a similarity score ranging from zero to six. A score of zero indicated no similarity or a deletion. The score for a pair of the same amino acid was typically five, but it was sometimes six for the amino acids considered less common by McLachlan (1972). These less common amino acids consisted of: phenylalanine, methionine, tyrosine, histidine, cysteine, tryptophan, arginine, and glycine (McLachlan, 1972). The score was normalised to give a scoring range of zero to ten, to avoid having two maximum scores.

Two approaches were used: 1) the 'fully-annotated' approach where the physicochemical similarity of the aligned residues was scored if both were annotated as catalytic (A in Figure 3.3), 2) the 'partially-annotated' approach where at least one residue should be annotated as being catalytic for an equivalent position to be scored (B and C in Figure 3.3). The latter approach allows for missing annotations or mis-annotations in the CSA. In both approaches, a catalytic residue aligned to a gapped position was penalised with the lowest score of zero (D and E in Figure 3.3). Scores were accumulated across the catalytic residue positions in an alignment and divided by the number of positions scored.

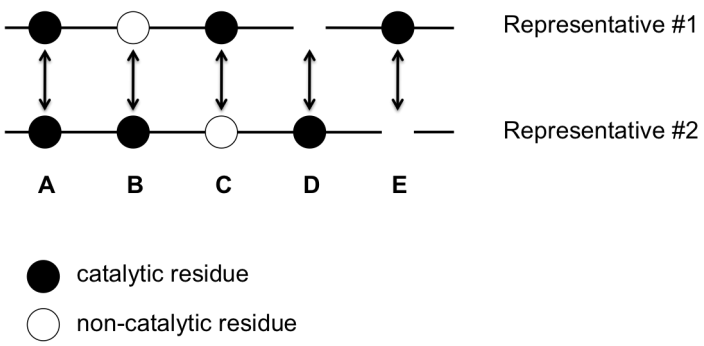


Figure 3.3: Similarities in residues at aligned positions are used to calculate the catalytic residue similarity between two functional family representative domain sequences.

3.2.2.2 Comparing catalytic machinery similarity using the 3D structure superposition method

There may be cases where the same catalytic residues in two related, superposed structural domains align well in 3D space, but do not align well at the sequence level. To identify these cases, all pairs of functional family representative domains were superposed within each superfamily and pairs of catalytic residues were identified that superposed within a distance of 5Å (see Figure 3.4).

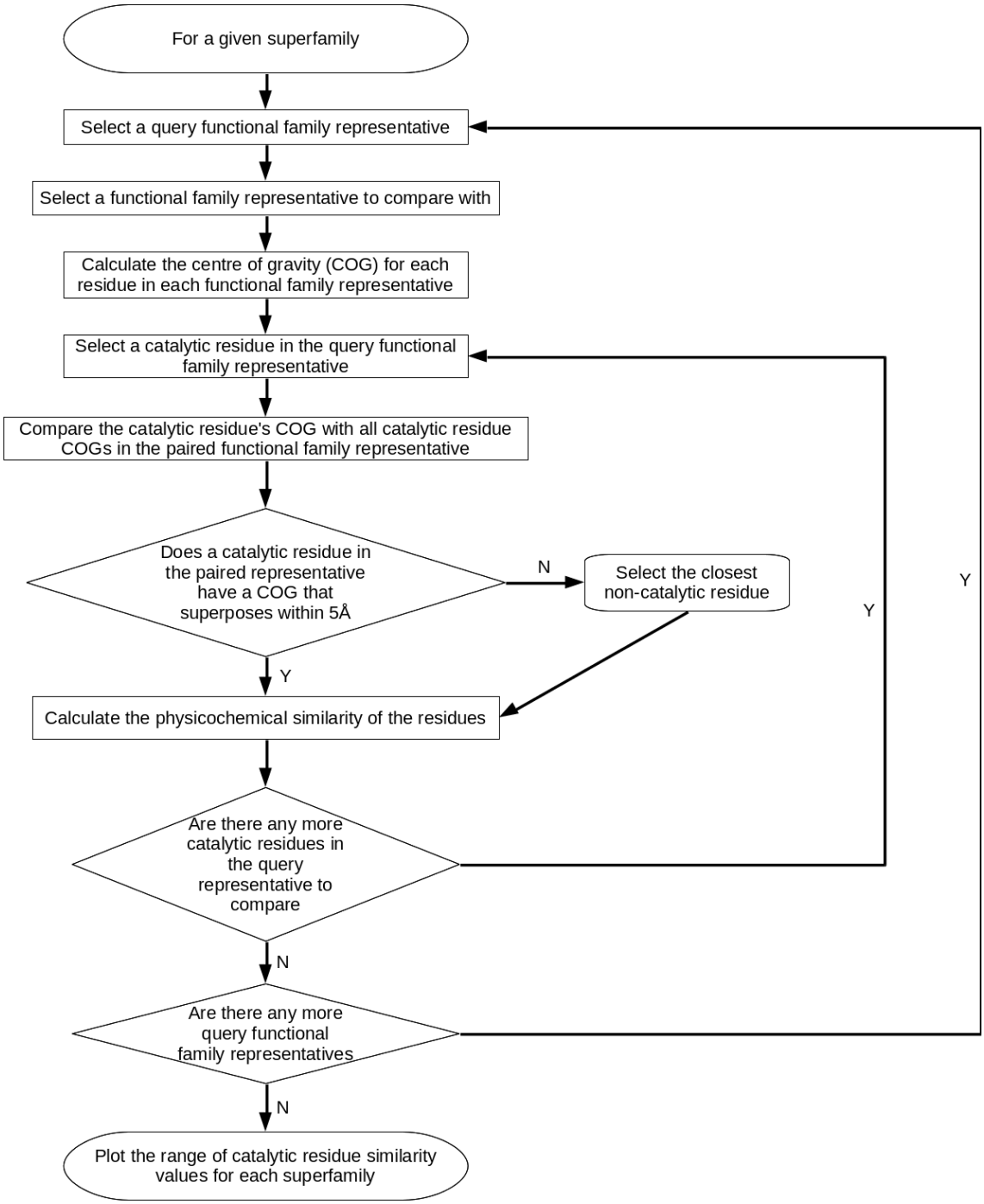


Figure 3.4: The steps taken when calculating catalytic residue similarity between a given pair of functional family representatives using the 3D structure superposition mapping protocol.

3.2.3 Comparing chemical reaction mechanisms

To compare the similarity of the chemical reaction mechanisms carried out across a superfamily, the EC-BLAST algorithm was used to automatically calculate the

similarity between all IUBMB enzyme reactions observed in the superfamily. These reactions were mapped from the EC numbers found within each superfamily. Enzyme reaction comparisons were made with respect to bond change, reaction centre, and small molecule substructure similarity scores (Rahman *et al.*, 2014).

As we only wanted to compare chemical reaction mechanism similarities between functional families which had a parent-child evolutionary relationship, information from FunTree (version 1.0) (discussed in Section 3.1.3.1) phylogenetic trees was used to describe the evolutionary relationships between FineFam functional families in a superfamily.

Comparisons of chemical reactions were done in collaboration with Nick Furnham in the Thornton Group at the EBI, who developed the FunTree resource.

3.2.4 Examining the structural preference of catalytic residues

To find the structural location of catalytic residues the secondary structure property of each catalytic residue was determined using the DSSP program. The eight categories from the DSSP program were reduced down to four by the BioPerl DSSP module (Stajich *et al.*, 2002): 1) 'H' represents helix secondary structure elements and incorporates alpha helix (H in DSSP), 3/10 helix (G), and Pi helix (I) states; 2) 'B' represents beta secondary structure elements and incorporates residues in isolated beta-bridges (B), and extended strands (E); 3) 'T' represents turns, which includes residues involved in H-bonded turns (T) and bends (S); 4) finally, '' represents no secondary structure assignment, i.e. loop regions.

The turn and no secondary structure assignment categories were merged into one and are referred to as loop regions in the chapter. The proportion of catalytic residues in each of the three categories was used to explore any preference in structural position.

3.3 Results

3.3.1 Identifying changes in catalytic machinery between functional families

Changes in catalytic residues, or catalytic machineries, within and between functional families have been identified and quantified within enzyme superfamilies. A total of 101 enzyme superfamilies were found to have catalytic site data present in two or more of their functional family representatives, and these superfamilies were therefore used in the analysis.

Before examining the changes in catalytic residues across functional families in our superfamily dataset, we first checked whether the FunFam$_{SEQ}$ functional families were sufficiently functionally coherent to perform this type of analysis.

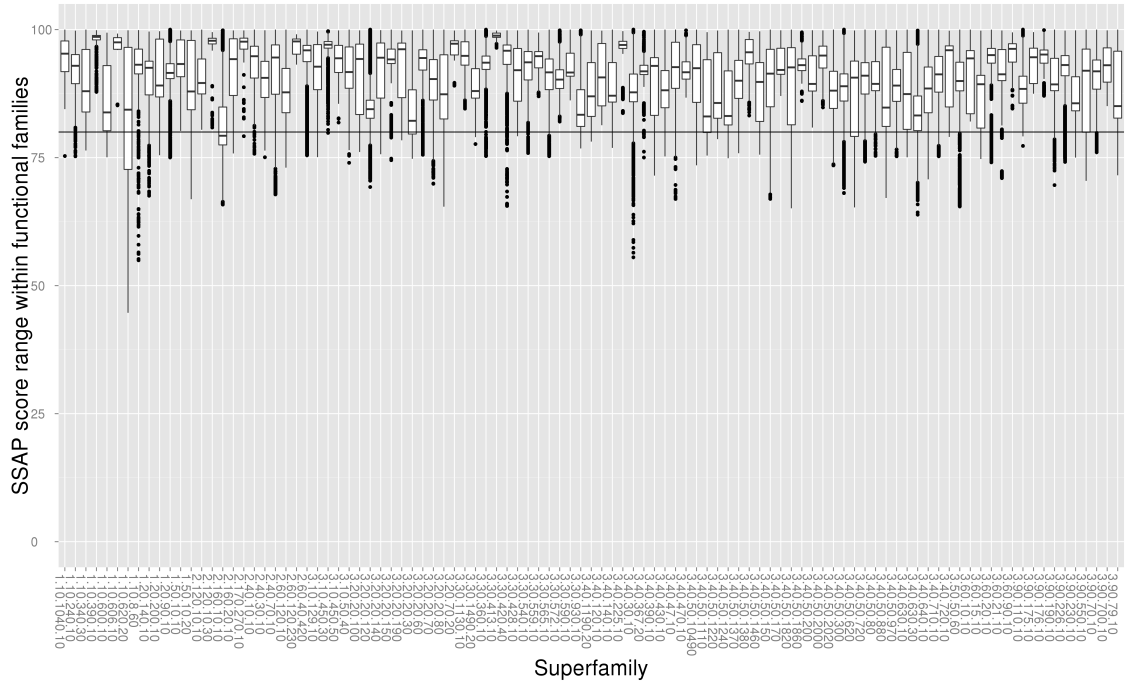
3.3.1.1 Examining the purity of the functional families

The previous chapter showed that the Funfamer method generated more functionally pure families that the DFX method. However, to examine enzyme superfamilies in more detail, the functional purity of FunFam$_{SEQ}$ functional families was also considered by examining the number of EC terms within them. Initial analysis showed that some FunFam$_{SEQ}$ functional families had multiple EC terms. Subsequently, the functional purity of each functional family was examined through the structural comparison of relatives, as well as through catalytic residue and EC number similarity.

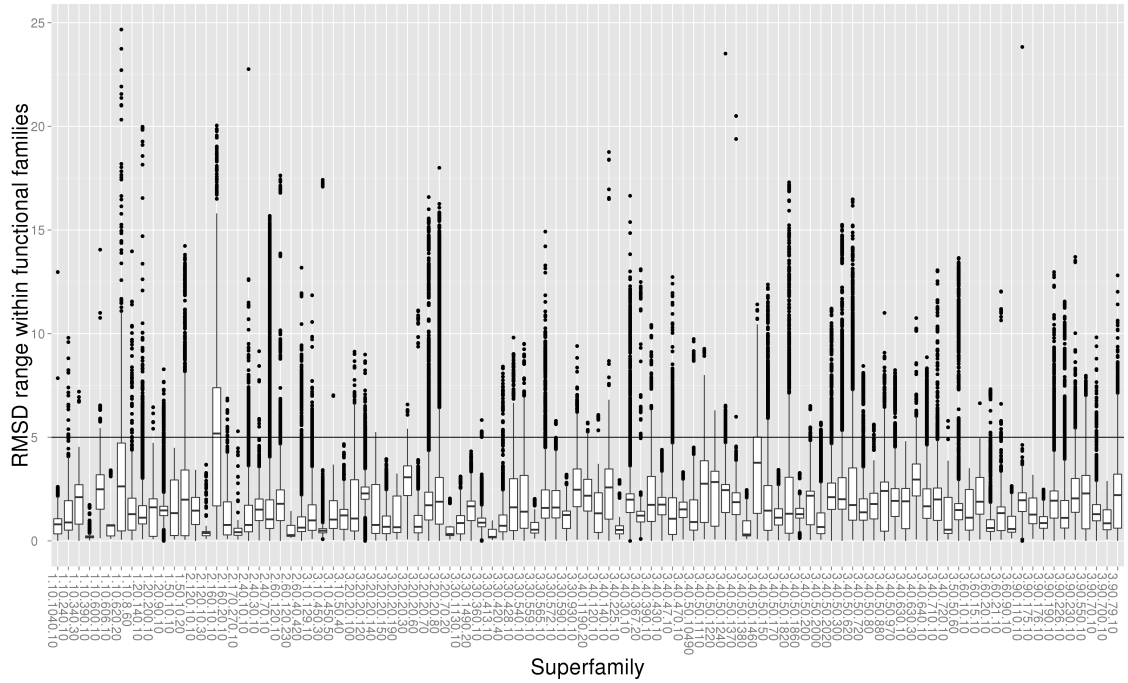
Testing functional coherence within functional families by measuring the structural similarity of relatives As FunFam$_{SEQ}$ functional families ideally contain sequences that code for a similar biological function, it was assumed that relatives would have high structural similarity. This is because previous studies have shown a correlation between structural and functional similarity (Redfern *et al.*, 2008; Dessailly *et al.*, 2010).

All domains within the functional families were pairwise structurally aligned with SSAP. A SSAP score of 80 or greater has been shown to indicate highly similar homologous structures (Orengo and Taylor, 1996) and functionally similar homologues tend to have an RMSD value of less than 3 Å (Cuff *et al.*, 2009).

Figure 3.5 shows that in the majority of the 101 superfamilies examined, the median values for the comparison of FunFam_{SEQ} functional family relatives are within the SSAP and RMSD thresholds. There are some outliers in a number of superfamilies.



(a) SSAP score measure.



(b) RMSD measure.

Figure 3.5: The structural comparison of FunFam$_{SEQ}$ functional family relatives in the enzyme dataset. Lines are drawn to highlight the points above a SSAP score of 80 and below an RMSD value of 5Å, which are acceptable thresholds for structural coherence. Functionally similar homologues tend to have an RMSD < 3Å.

The 'Ribonucleotide Reductase Subunit A' CATH superfamily (ID: 1.10.620.20)

is an example of a superfamily that contains both structurally coherent and structurally diverse FunFam$_{SEQ}$ functional families. It contains three FunFam$_{SEQ}$ functional families. While all of the domains in one family (ID: 1808 in terracotta) (Figure 3.6) and most of the domains in a second family (ID: 1818 in blue) are highly structurally similar, many of the domains in the third family (ID: 1817 in green) are structurally diverse. The presence of outliers in some FunFam$_{SEQ}$ functional families suggested that these families were not sufficiently structurally and functionally pure, and should be subclassified for our study of changes in catalytic residues across superfamilies.

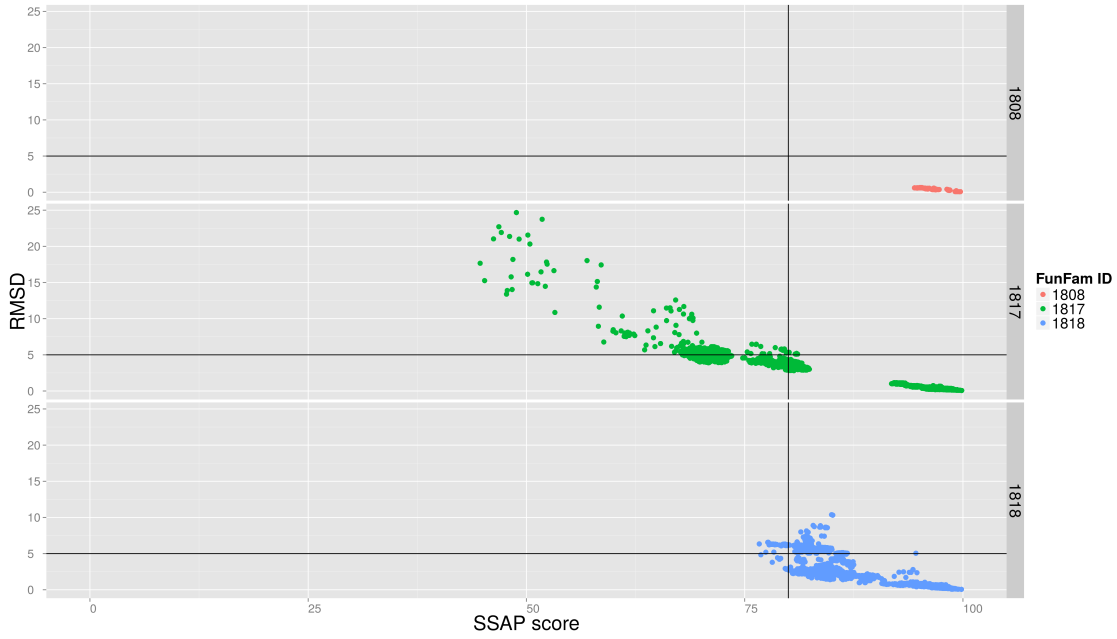
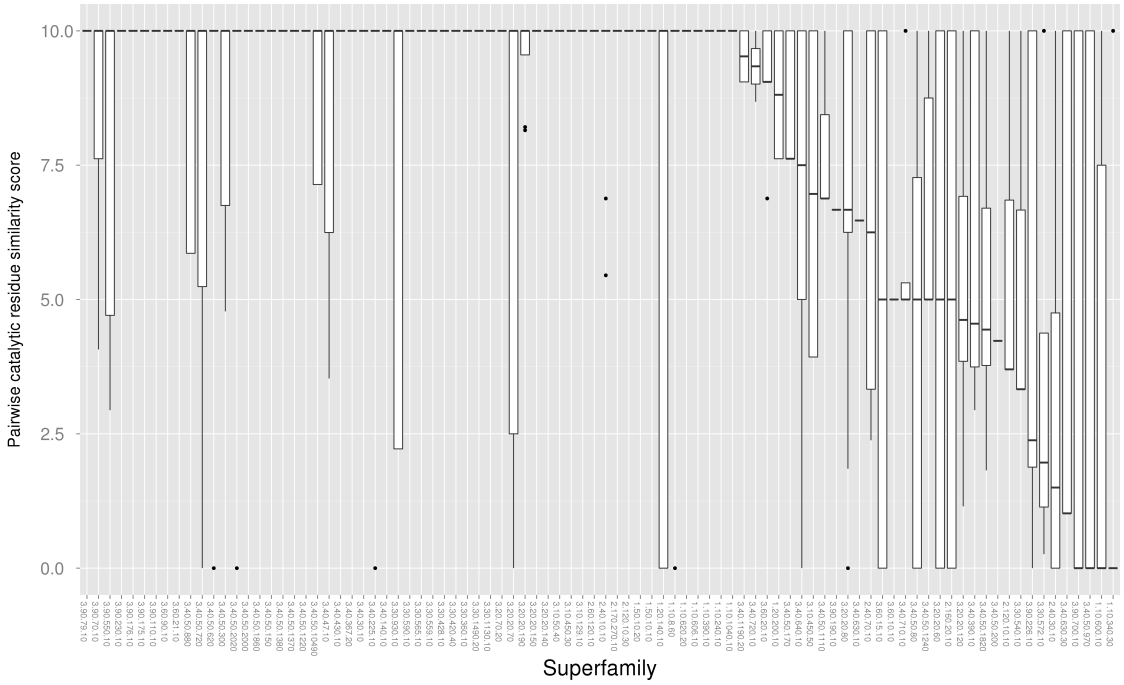


Figure 3.6: The level of structural diversity found within FunFam$_{SEQ}$ functional families in the superfamily 1.10.620.20. The contours represent the areas of highest density and the black lines represent the RMSD and SSAP score thresholds.

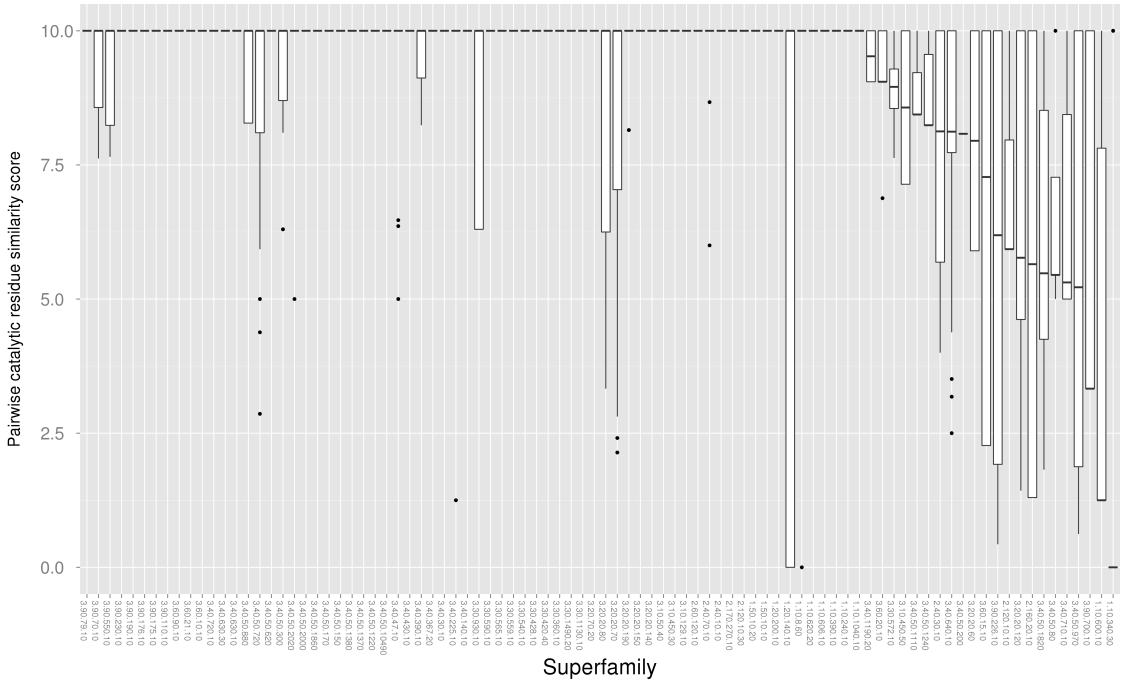
Testing functional coherence within a functional family by measuring catalytic residue similarity Another way of finding out whether FunFam$_{SEQ}$ functional families are functionally coherent is to look at the conservation of functional site residues across the family. The method described in Section 3.2.2 was used to compare the catalytic residue similarity.

An RMSD filter of 5Å was applied to ensure that the sequence alignment was

based upon structures that were not too divergent from each other. Empirical analyses by CATH curators have shown at this RMSD it is possible to get accurate structural alignments of homologues. Figure 3.7 shows that even when this RMSD filter was applied, there are a few families identified with highly variable catalytic residues based on both approaches. 33 (36.67%) and 22 (24.44%) of superfamilies using the fully-annotated and the partially-annotated approaches (see definition in Methods Section 3.2.2.1, page 117), respectively, have a median similarity value less than ten.



(a) Scoring aligned pairs using the fully-annotated approach.



(b) Scoring aligned pairs of residues using the partially-annotated approach.

Figure 3.7: The similarity of catalytic residues within FunFam$_{SEQ}$ functional family representatives. Catalytic residues were mapped using pairwise structure-based sequence alignment information. The data are ranked by median similarity score.

3.3.1.2 Subclassifying the functional families into subfamilies with greater functional coherence

To improve their functional coherence, the FunFam$_{SEQ}$ functional families from the 101 enzyme superfamily data set were subclassified on the basis on their EC annotations, so that a single EC term was associated with each subfamily. Up to five functional families per superfamily were split using EC information. 29 of the 101 superfamilies had at least one functional family split, and 9 of the 101 superfamilies had two or more functional families split. The percentage of families split are shown in Figure 3.8. These split families will now be referred to as FineFams (see Figure A.1).

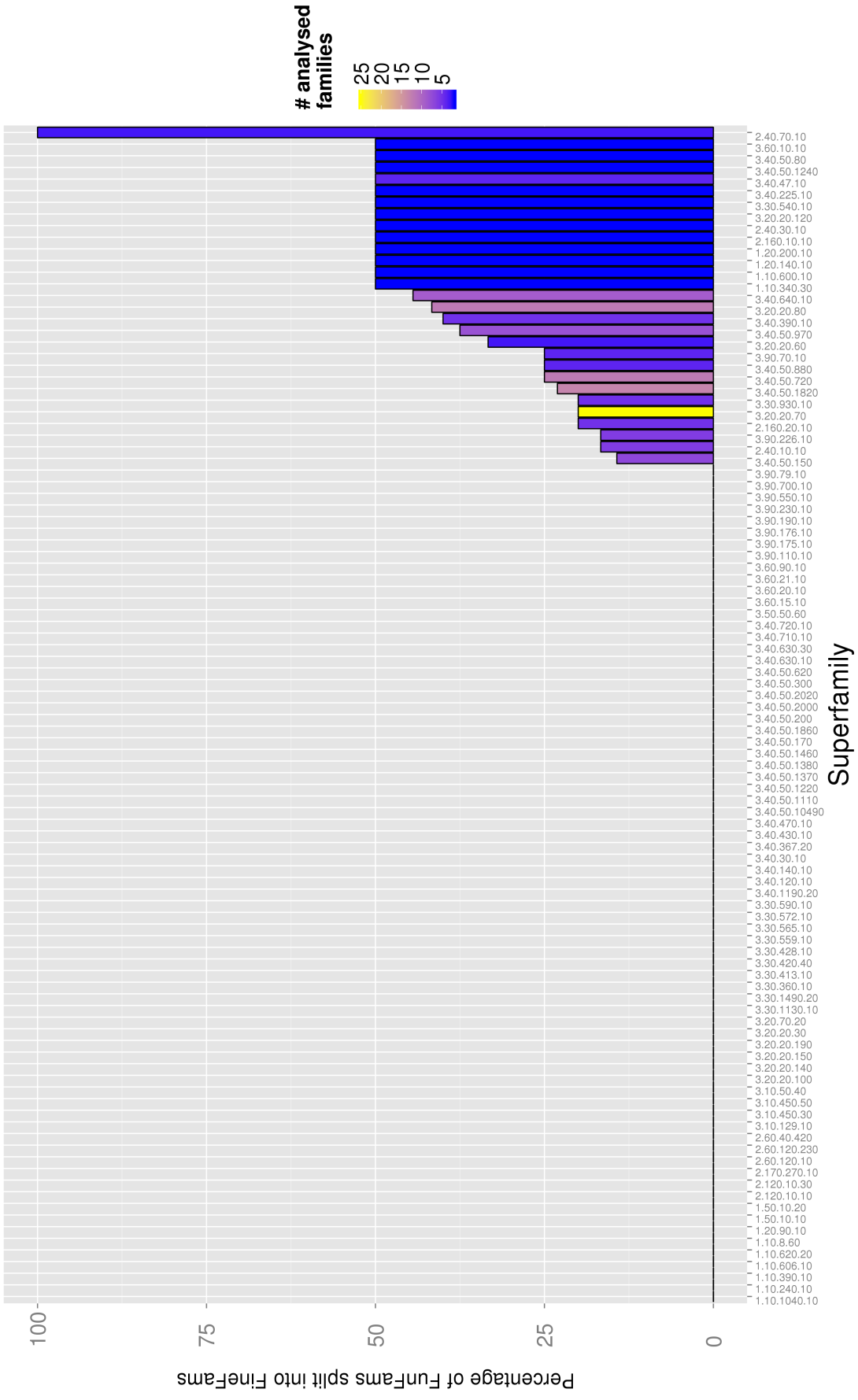
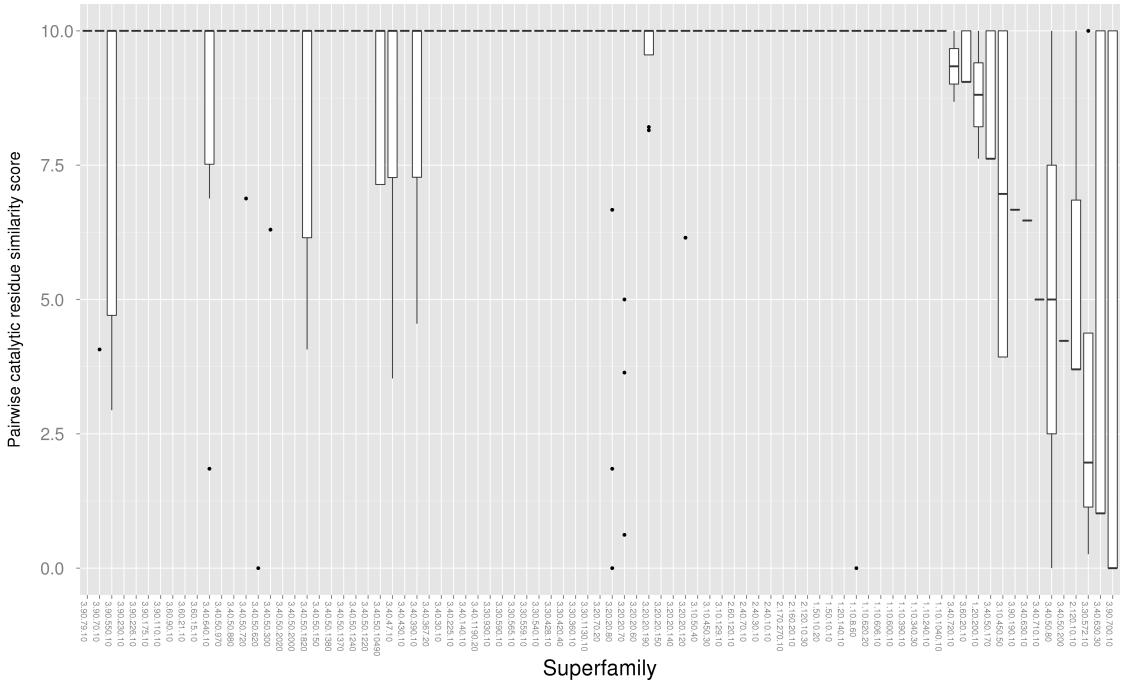
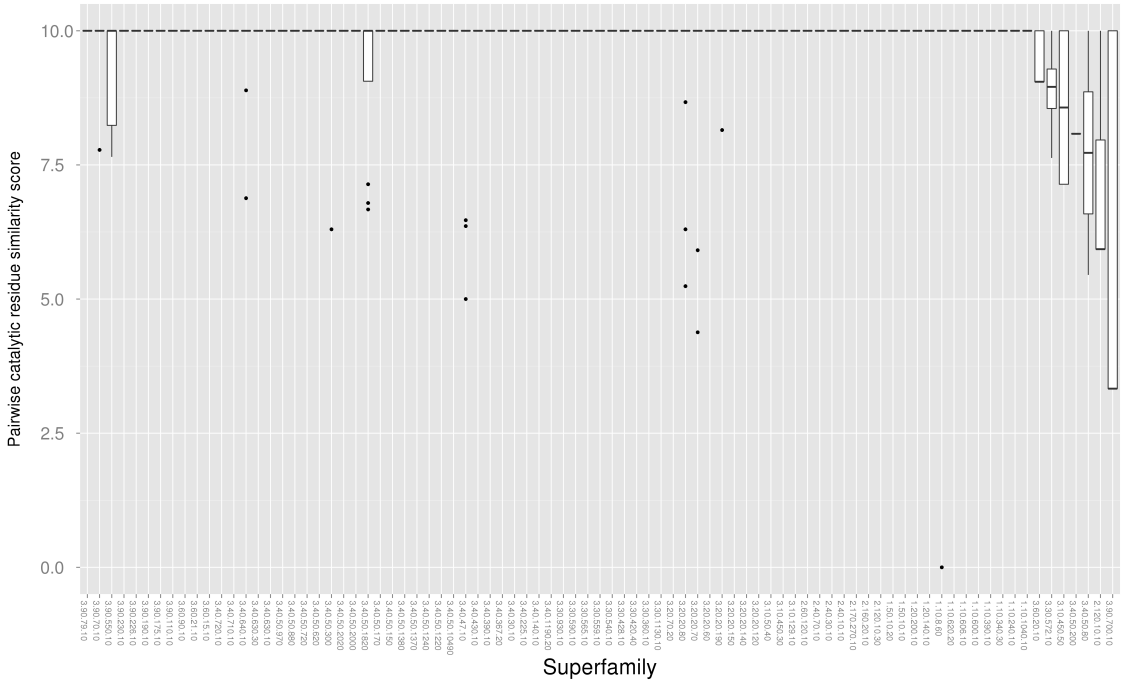


Figure 3.8: The percentage of FunFam$_{SEQ}$ functional families in a superfamily containing different EC terms, which were split into FineFams. The colour scale represents the number of FunFam functional families in a superfamily with CSA residue information.

To explore whether the catalytic residue similarities had increased within functional families following splitting into FineFams, both the 'fully-annotated' and 'partially-annotated' approaches were used (see Figure 3.9). In 85 superfamilies, FineFam functional families relatives could be compared as they superposed within 5Å. Using the fully-annotated approach, 71 superfamilies (83.53%) have pairs of family representatives with a median similarity score of 10 (see Figure 3.9a). Using the partially-annotated approach, 78 superfamilies (91.76%) have family pairs with a median score of 10. These results show a very encouraging increase in catalytic residue similarity from the FunFams to the FineFams. Approximately 15-20% more superfamilies have FineFam functional family relatives that align well and have a median similarity score of ten.



(a) Scoring aligned pairs using the fully-annotated approach.



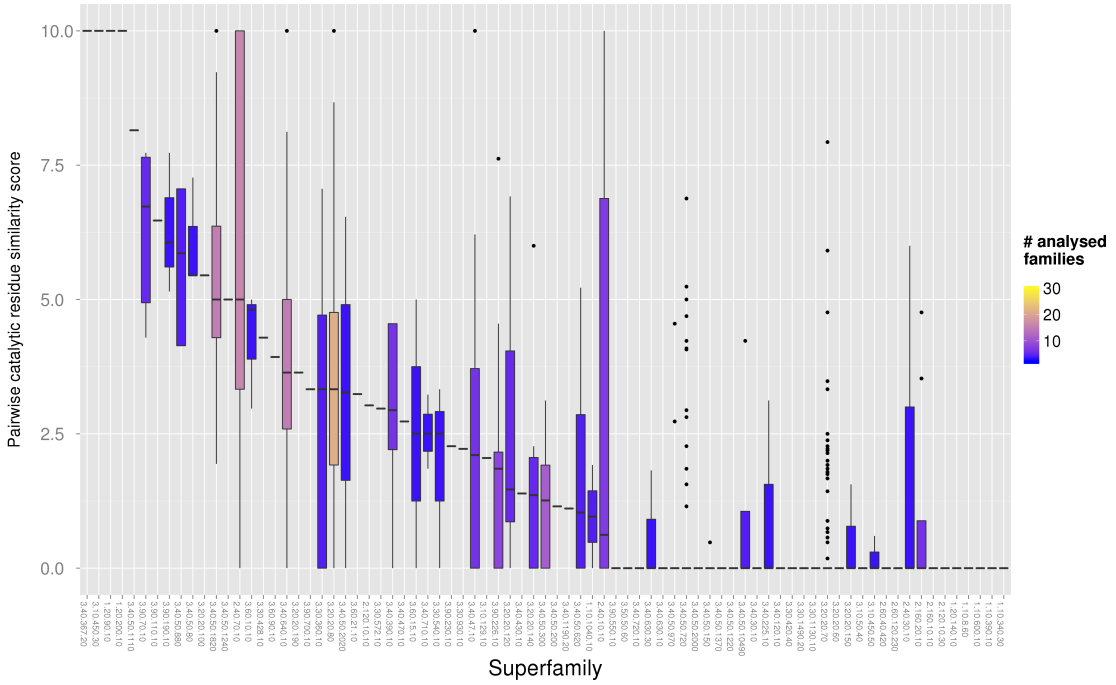
(b) Scoring aligned pairs of residues using the partially-annotated approach.

Figure 3.9: Catalytic residue similarity within FineFam functional families. Catalytic residues were mapped using pairwise structure-based sequence alignment information.

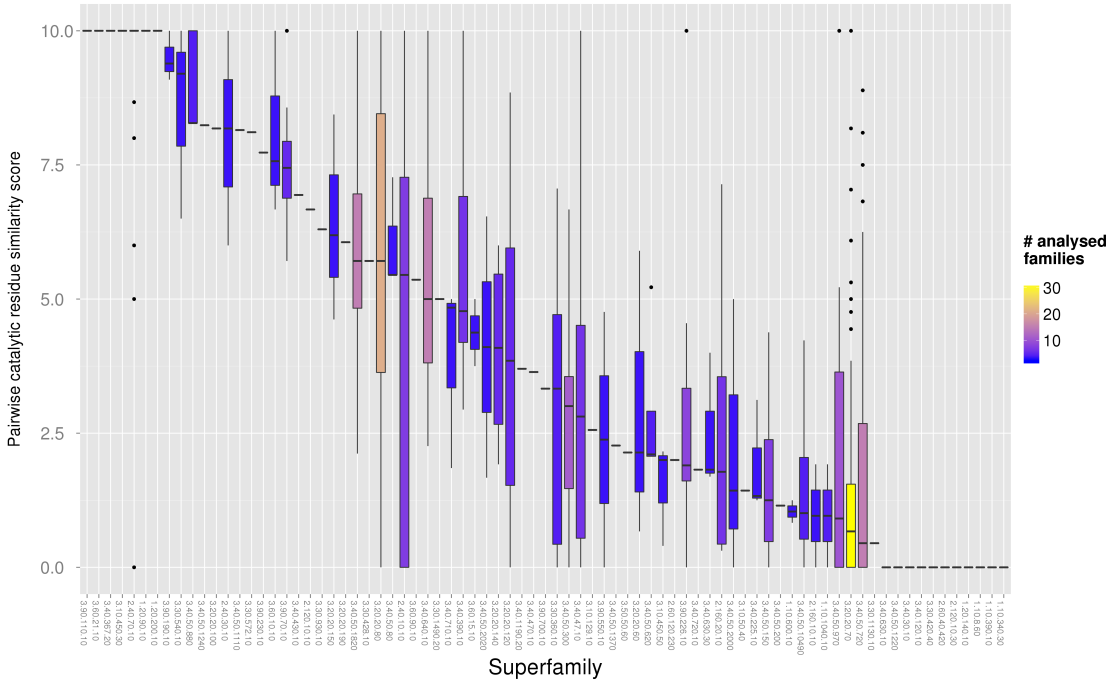
3.3.1.3 Comparing catalytic machinery similarity between FineFam functional families across superfamilies

To explore the extent to which catalytic machineries can change across superfamilies, catalytic residues were compared between FineFam functional families across all superfamilies in the 101 dataset. The catalytic residue alignment information, for all the superfamilies, has been made available online at `http://www.cathdb.info/nataliedawson/csashift/list_all_superfamilies`. However for the analyses described below, superfamilies were only included in the analyses if all their functional family representatives could be superposed within 5Å. This gave a dataset of 79 superfamilies.

a. Mapping catalytic residues between functional families using the pairwise structure-based sequence alignment protocol Surprisingly, only 14 (17.72%) superfamilies have a median similarity score of at least 5 using the fully-annotated approach and 31 (39.24%) superfamilies using the partially-annotated approach (see Figure 3.10). Figure 3.10 reports a wide range of different catalytic machineries used between relatives for the majority of superfamilies analysed. In addition, there is no clear correlation between the number of FineFam functional families analysed in a superfamily and the catalytic residue similarity score.



(a) Scoring aligned pairs using the fully-annotated approach.

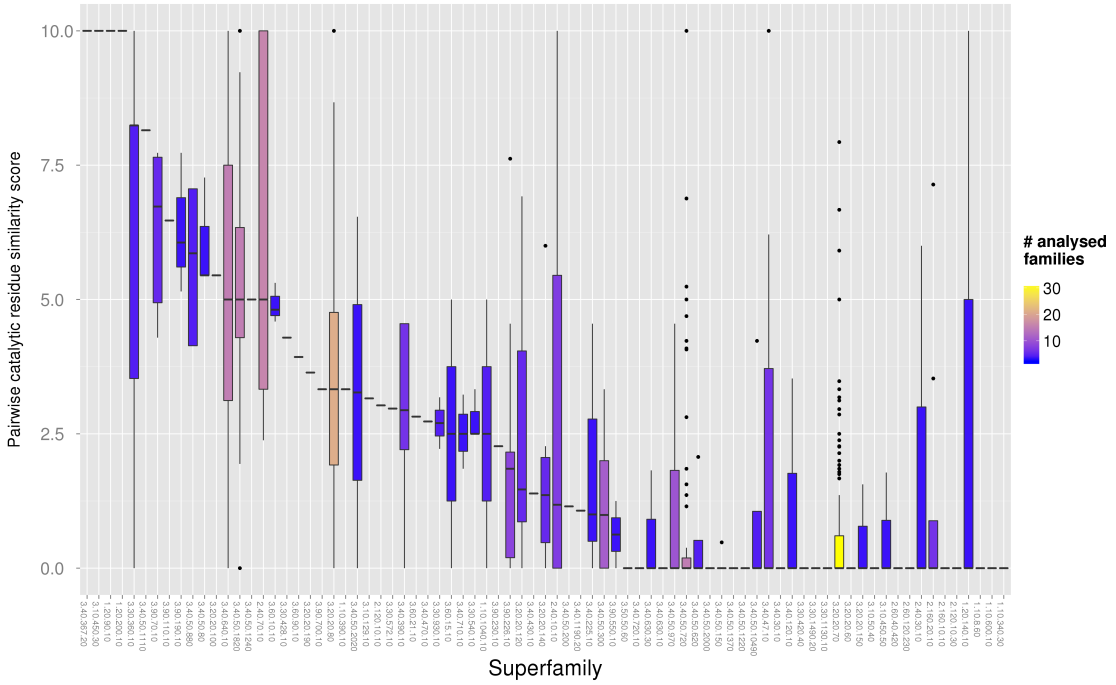


(b) Scoring aligned pairs of residues using the partially-annotated approach.

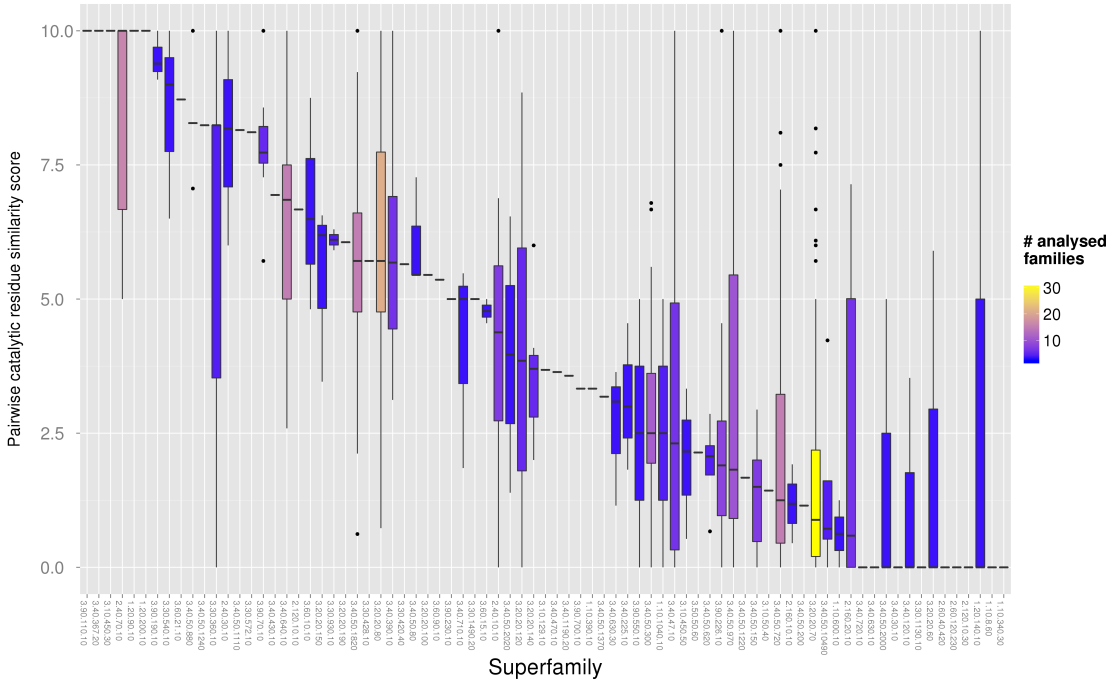
Figure 3.10: Catalytic residue similarity between FineFam functional family representatives. Catalytic residues were mapped using the pairwise structure-based sequence alignment protocol.

b. Mapping catalytic residues between functional families using the 3D superposition protocol If one uses spatial information to identify equivalent cat-

alytic residues between functional families, a similar range of similarity scores are observed between FineFams (see Figure 3.11). In only 16 (20.25%) and 34 (43.04%) superfamilies, using the fully-annotated and partially-annotated approaches, respectively, a similarity score of at least five is reported. For the majority of families there is considerable change in the catalytic residues in the active site.



(a) Scoring aligned pairs using the fully-annotated approach.



(b) Scoring aligned pairs of residues using the partially-annotated approach.

Figure 3.11: Catalytic residue similarity between FineFam functional family representatives. Catalytic residues were mapped using the 3D superposition-based protocol.

Figure 3.12 compares the similarity scores for each FineFam pair obtained using either the pairwise sequence alignment based protocol to identify equivalent catalytic

residues (x-axis) or the 3D structure superposition based protocol (y-axis). The 'partially-annotated' approach was used to assess catalytic residue similarities.

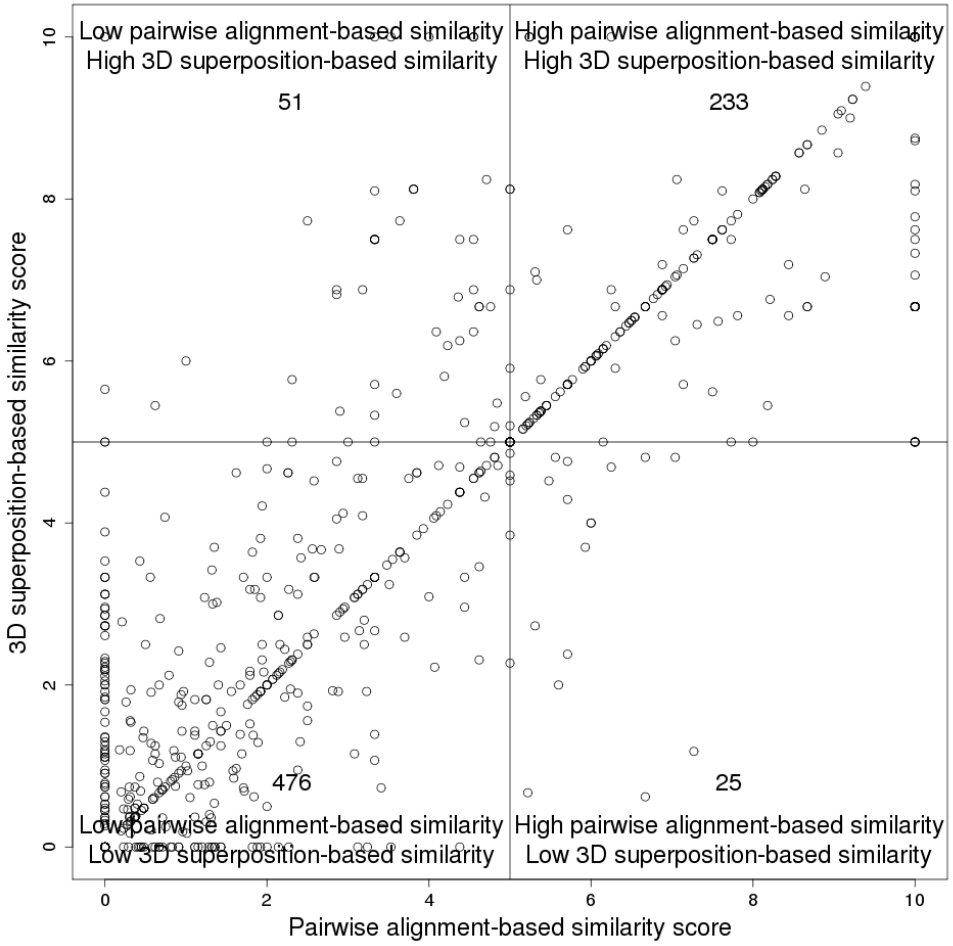


Figure 3.12: Comparing the pairwise alignment-based catalytic residue similarity score against the 3D superposition-based catalytic residue similarity score for each pair of FineFam representative domains. The number of pairs compared is noted in each quadrant.

Figure 3.12 shows that there is a correlation between the schemes' scores for some functional families and that the most populated quadrants are those representing low pairwise sequence alignment- and low 3D structure superposition-based scores, and high pairwise sequence alignment- and high 3D structure superposition-based scores.

Unsurprisingly, a large number (233, 29.68%) of FineFam pairs have a high pairwise alignment-based similarity score accompanied with a high 3D superposition-

based similarity score. As would also be expected, a correlation is seen between low pairwise alignment-based and low 3D superposition-based scores and the majority of pairs (476, 60.64%) are in this category. A Pearson's product-moment correlation of 0.872 was calculated between all pairs of similarity scores.

Despite this high correlation between the two types of similarity score, variation can be observed between some pairs. For example, 51 FineFam pairs have catalytic residues that do not align well when using a pairwise structure-based sequence alignment but that closely co-locate in 3D. This may suggest that relatives in different functional families have evolved to perform similar functions using the same catalytic residues, but these residues are contributed from different positions in the sequence.

A pair of functional family representatives from the 'Butyryl-CoA Dehydrogenase, subunit A, domain 3' CATH superfamily (ID: 1.20.140.10) are an example of two functional families having a catalytic residue similarity scores of 0 and 10 depending on whether they are mapped by the pairwise sequence alignment or the 3D superposition protocol, respectively. Figure 3.13 illustrates the superpositions of the two representatives and highlights a catalytic glutamic acid residue, which is not found at the same position in the sequence (see Figure 3.14) but is co-located to within 5Å in the two domain structures.

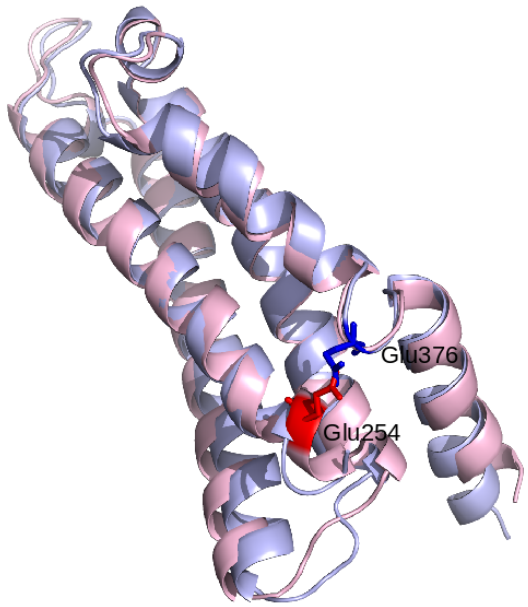


Figure 3.13: Two functional family representative domains whose catalytic residues superpose within 5Å but that do not align at the sequence level. The domains represent subunits of isovaleryl-CoA dehydrogenase (CATH domain ID: 1ivhA03 in light pink) and acyl-CoA dehydrogenase (CATH domain ID: 3mddA03 in light blue).

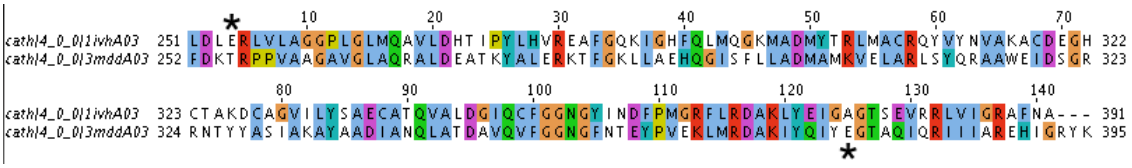
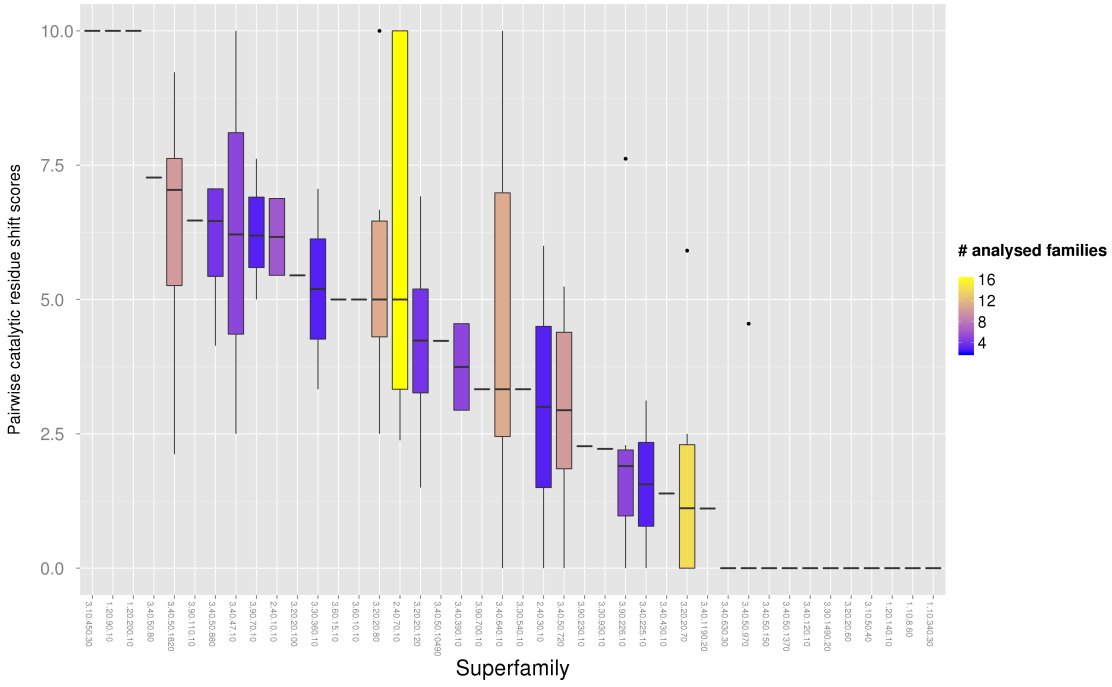


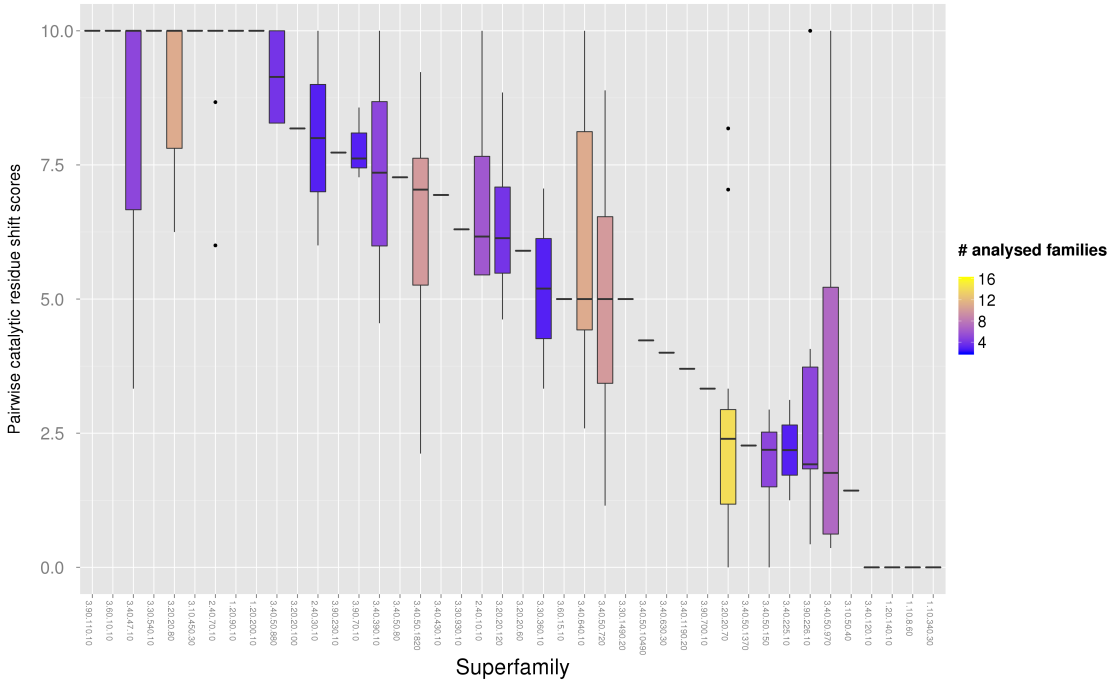
Figure 3.14: Pairwise structure-based sequence alignment with SSAP of the two functional family representatives, 1ivhA03 and 3mddA03, visualised in Jalview (Waterhouse *et al.*, 2009). The single glutamic acid residue used by both domains does not align and can be found at opposite ends of the alignment (see asterisks).

In the bottom-right quadrant of Figure 3.12, a small number (25) of FineFam pairs are reported to have high similarity using the pairwise alignment protocol and a low similarity using the 3D superposition protocol. This may suggest that some relatives in different functional families have structural embellishments that result in the catalytic residues being positioned in different spatial position despite being in a similar sequence position.

When comparing the conservation of functional residues between two closely related protein domains via a structure-based sequence alignment, there is no benchmarked RMSD value that reflects sufficient structural similarity to ensure an accurate alignment. In order to ensure that our protocol gave an accurate alignment of catalytic residues we have only compared functional family representatives that superpose within an RMSD of 5 Å. However even at this threshold relatives may be too diverse, and this could be the cause of the wide range of catalytic residue similarities reported across superfamilies. A cut-off RMSD of 3Å was therefore also used to examine whether such variation is observed between more structurally-similar domains. Figure 3.15 shows that the range of catalytic residue similarity is still just as varied when using the lower RMSD threshold, i.e. a significant proportion of superfamilies are still showing low similarity in catalytic residue machinery between functional families.



(a) Scoring aligned pairs using the fully-annotated approach.



(b) Scoring aligned pairs of residues using the partially-annotated approach.

Figure 3.15: The catalytic residue similarity between FineFam functional family representatives, which superpose within 3Å.

3.3.2 Exploring whether different catalytic machineries used within enzyme superfamilies are associated with different enzyme chemistries

In the last section, catalytic residue similarity scores were reported across 79 superfamilies. Of these superfamilies, 72 have FineFam functional families which use different catalytic machineries (see Figure 3.16) where different catalytic machineries are defined by having a catalytic residue similarity score less than five. We examined whether changes in catalytic machinery were associated with changes in enzyme chemistry, which were measured using EC terms. If the relatives have the same EC number at the third hierarchical level they were assumed to have the same enzyme chemistry.

Figure 3.16 shows the number of functional families that use different catalytic machineries in each of the 72 superfamilies. These superfamilies are divided according to the number of functional families that perform: 1) different enzyme chemistries, 2) the same enzyme chemistry but act on different substrates, or 3) the same enzyme chemistry and act on the same substrate.

In 50 superfamilies there are between 2 and 30 functional families where changes in catalytic machinery are associated with changes in enzyme chemistry (see red in Figure 3.16). Not reported previously on this scale, there are 54 superfamilies with functional families using different catalytic machineries to perform the same enzyme chemistry. Of these 54, 38 superfamilies have between 2 and 18 functional families that act on different substrates, i.e. they differ at fourth EC hierarchical level (see blue in Figure 3.16). While the residues from the CSA do not contain substrate-binding residues, some of the CSA residues could be involved in roles associated with the different substrates, for example stabilising the different intermediates. Interestingly, the remaining 16 of these 54 superfamilies each contain two or three functional families that use different catalytic machineries to perform the same enzyme chemistry with the same substrate (see yellow in Figure 3.16).

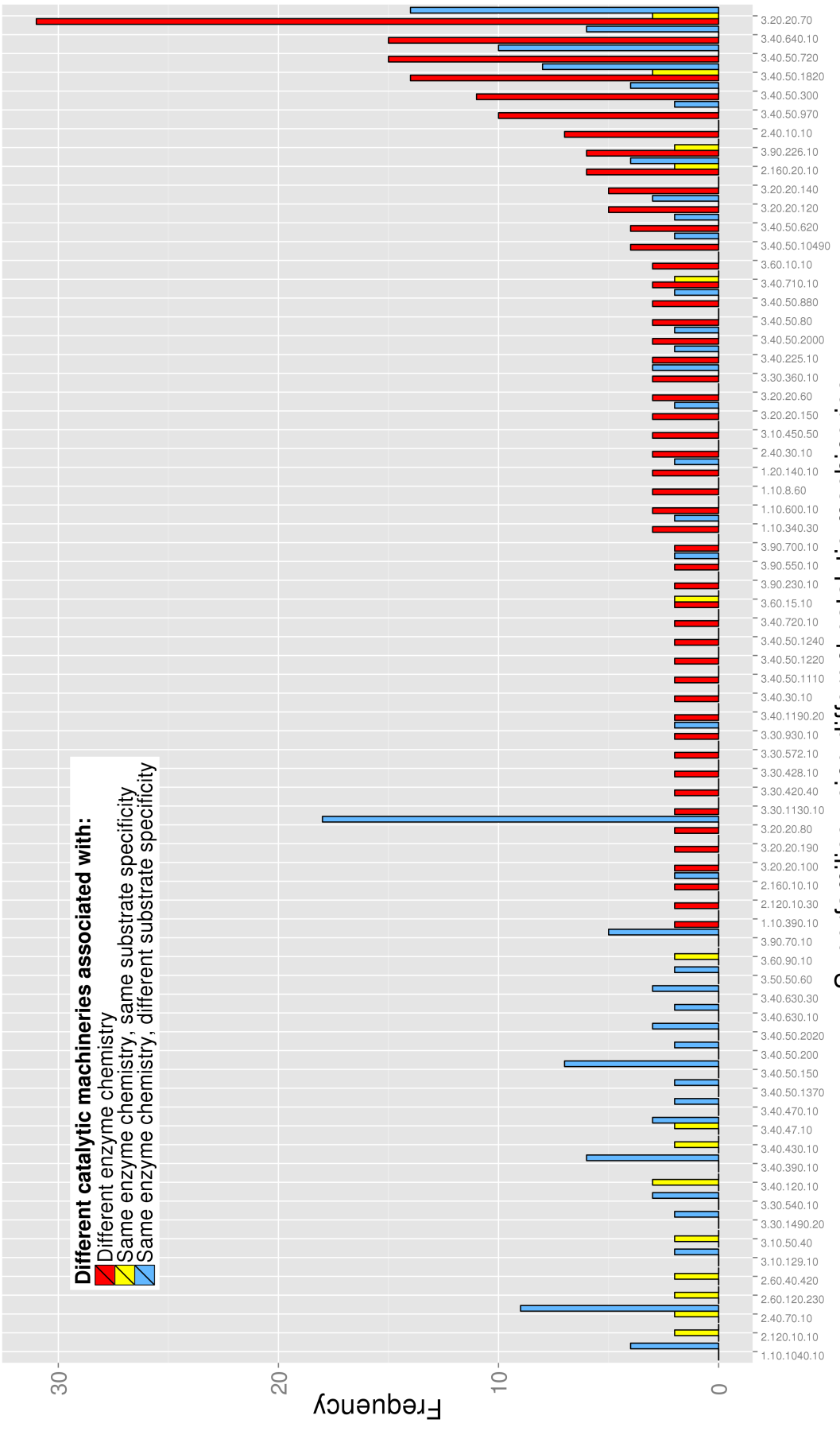


Figure 3.16: The number of FineFam functional families in each superfamily that use different catalytic machinery to: 1) catalyse an enzymatic reaction with different enzyme chemistries (in red), 2) catalyse an enzymatic reaction with the same enzyme chemistry but different substrates (in blue), and 3) catalyse an enzymatic reaction with the same enzyme chemistry and the same substrate (in yellow).

An example of two relatives that use different catalytic machineries to perform the same enzyme chemistry comes from the Thiamine diphosphate (TPP)-dependent domain superfamily (CATH ID: 3.40.50.970). Figure 3.17 shows the superposition of two TPP domains from pyruvate decarboxylase (PDC) and benzoylformate decarboxylase (BFD), which are both carboxy-lyase enzymes but they use different catalytic residues. Carboxy-lyases, also known as decarboxylases, catalyse the addition or the removal of a carboxyl group to or from a compound.

Both domains have a common histidine. In BFD, Glu28 is reported to possibly form a protein relay system with His70, through hydrogen bond formation (Polovnikova *et al.*, 2003) and as PDC also has a negatively-charged (Asp28) and positively-charged (His114 / His115), these residues could also form a protein relay system. Polovnikova *et al.* (2003) reported that, apart from the residues bound to the TPP cofactor, there is a general lack of conserved residues in the active site of TPP-dependent enzymes. A number of studies have reported however, that the histidine catalytic residues present in both PDC and BFD are positioned similarly with respect to the TPP cofactor (Arjunan *et al.*, 1996; Dyda *et al.*, 1993; Hasson *et al.*, 1998). Figure 3.18 illustrates the sequence alignment and highlights the differences in catalytic residue sequence positions between the two domains.

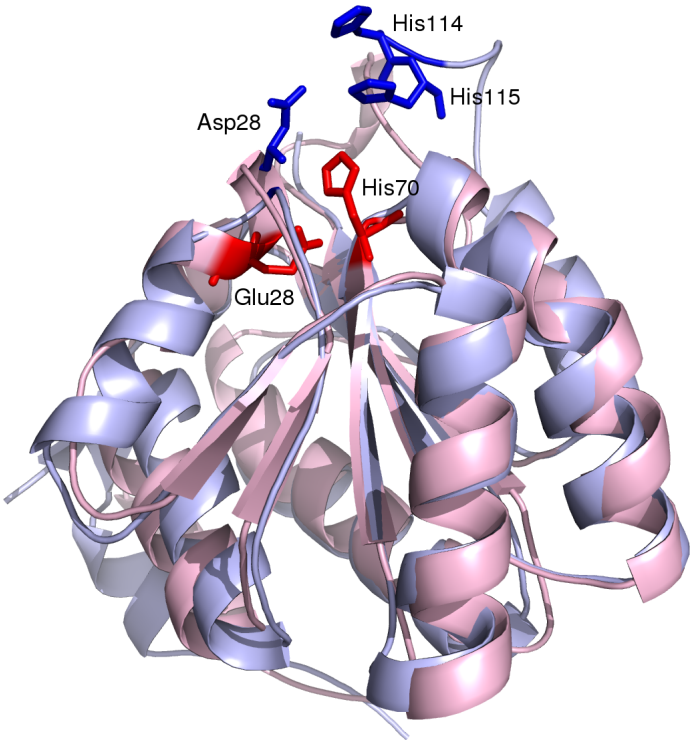


Figure 3.17: Two functional families from the Thiamine diphosphate (TPP)-dependent domain superfamily (CATH ID: 3.40.50.970) use different catalytic machineries to perform the same carboxylyase enzyme chemistry. Domains 1pvdA01 (in light blue) and 1bfdA02 (in light pink) are from the enzymes pyruvate decarboxylase (EC 4.1.1.1) and benzoylformate decarboxylase (EC 4.1.1.7), respectively.

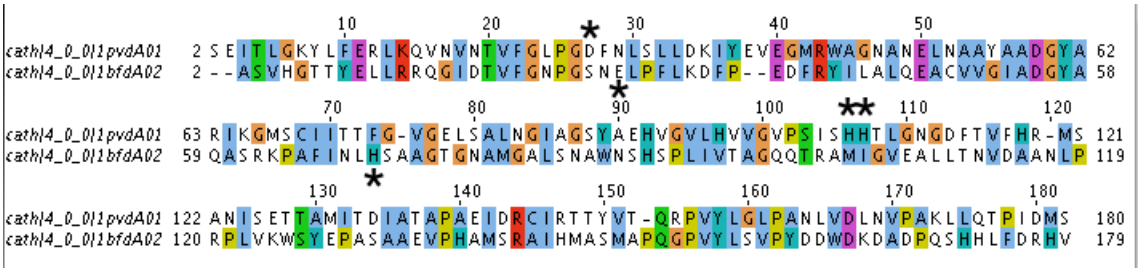


Figure 3.18: Pairwise structure-based sequence alignment with SSAP of the two functional family representatives, 1pvdA01 and 1bfdA02, visualised in Jalview (Waterhouse *et al.*, 2009). The two domains use different catalytic residues that do not align at the sequence level (see asterisks), however the domains perform the same reaction mechanism.

Clearly, two functional family representatives with a catalytic residue similarity score less than 5 still have some similarity in catalytic machinery, as shown here with the common histidine residue and the common negatively charged residue. Similarly, the 37 other superfamilies with some changes in catalytic residues and acting on

different substrates will also show some similarity in catalytic residues in terms of physicochemical properties conservation or residue location conservation. This is expected in agreement with the early work of Babbitt and Gerlt (1997) and Todd *et al.* (2001), who showed that substrate specificity is more likely to change than chemistry across a superfamily. They also showed that catalytic residues must be conserved within the active site to enable the common reaction steps to be preserved across the superfamily.

3.3.3 Examining the correlation between catalytic machinery and reaction mechanism across a superfamily

To explore these concepts further, we examined how frequently a change in catalytic machinery across a superfamily was accompanied by a change in the enzyme chemical reaction mechanism. The catalytic residue similarity score between pairs of FineFam functional families, measured using the McLachlan scoring matrix (see Methods Section 3.2.2.1, page 119), was used to report a change in catalytic machineries, and the reaction mechanism similarity score, was measured using EC-BLAST (see Methods Section 3.2.3, page 119). To ensure the accuracy of the study, only functional families which had been assigned SwissProt-curated EC4 terms were used.

Figures 3.19, 3.20, and 3.21 plot the catalytic residue similarity score and the reaction mechanism similarity score between pairs of functional family representatives. While there is no clear correlation between these two properties there are a number of functional family pairs (shown in the top-right quadrants) that perform the same reaction mechanisms using the same catalytic machineries. Also, unsurprisingly, there is a greater density of points in each of the bottom-left quadrants, i.e. changes in catalytic residues accompanied by differences in bond changes, reaction centres, and substructures.

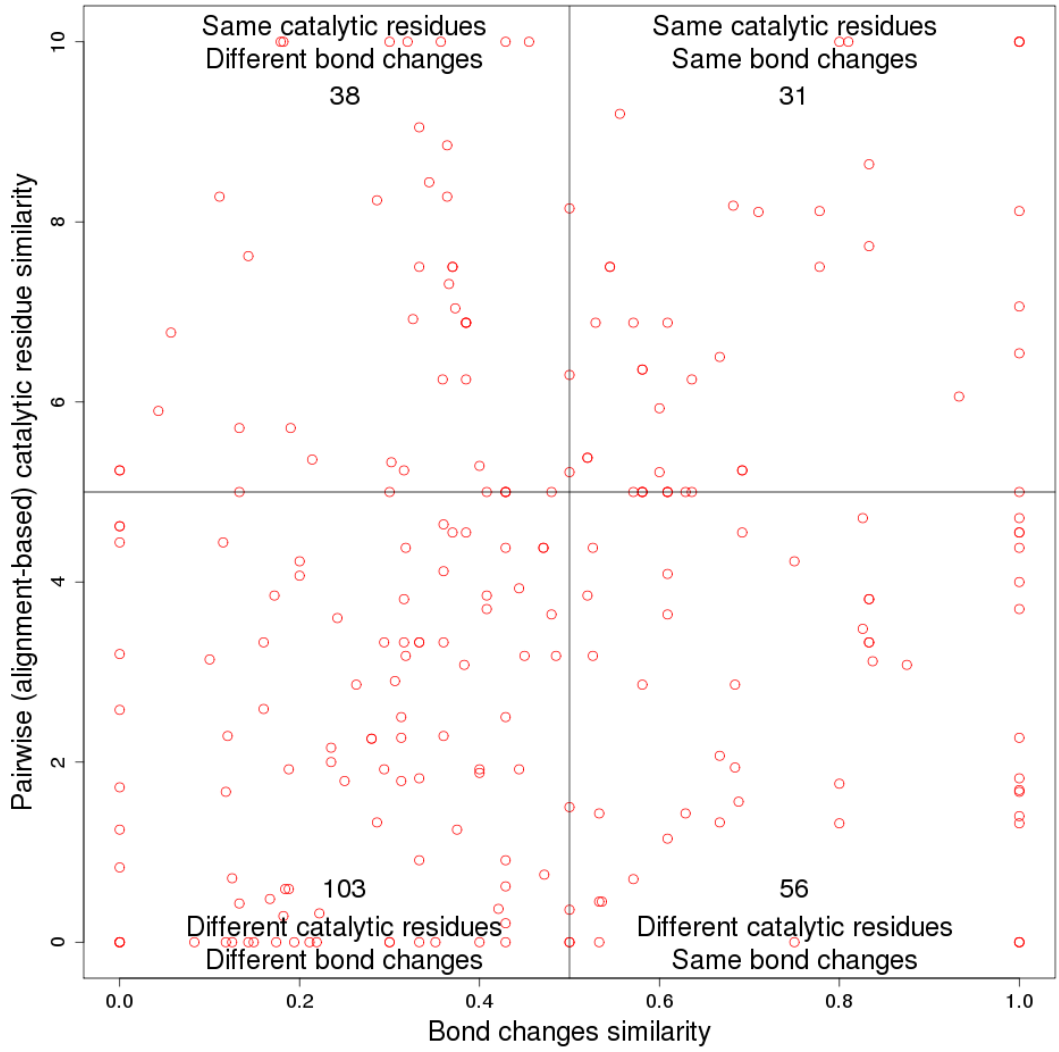


Figure 3.19: Correlating catalytic residue similarity with similarity in bond change.

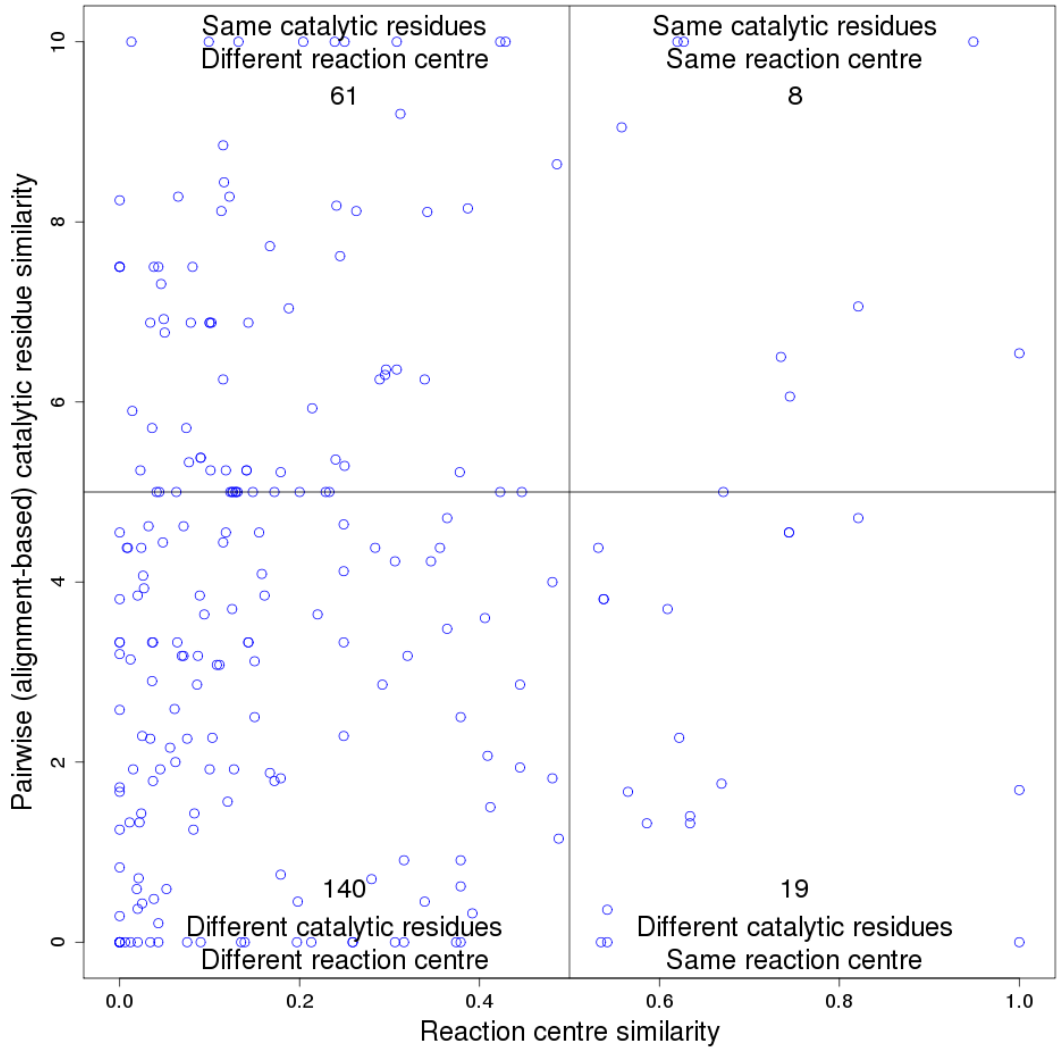


Figure 3.20: Correlating catalytic residue similarity with similarity in reaction centre.

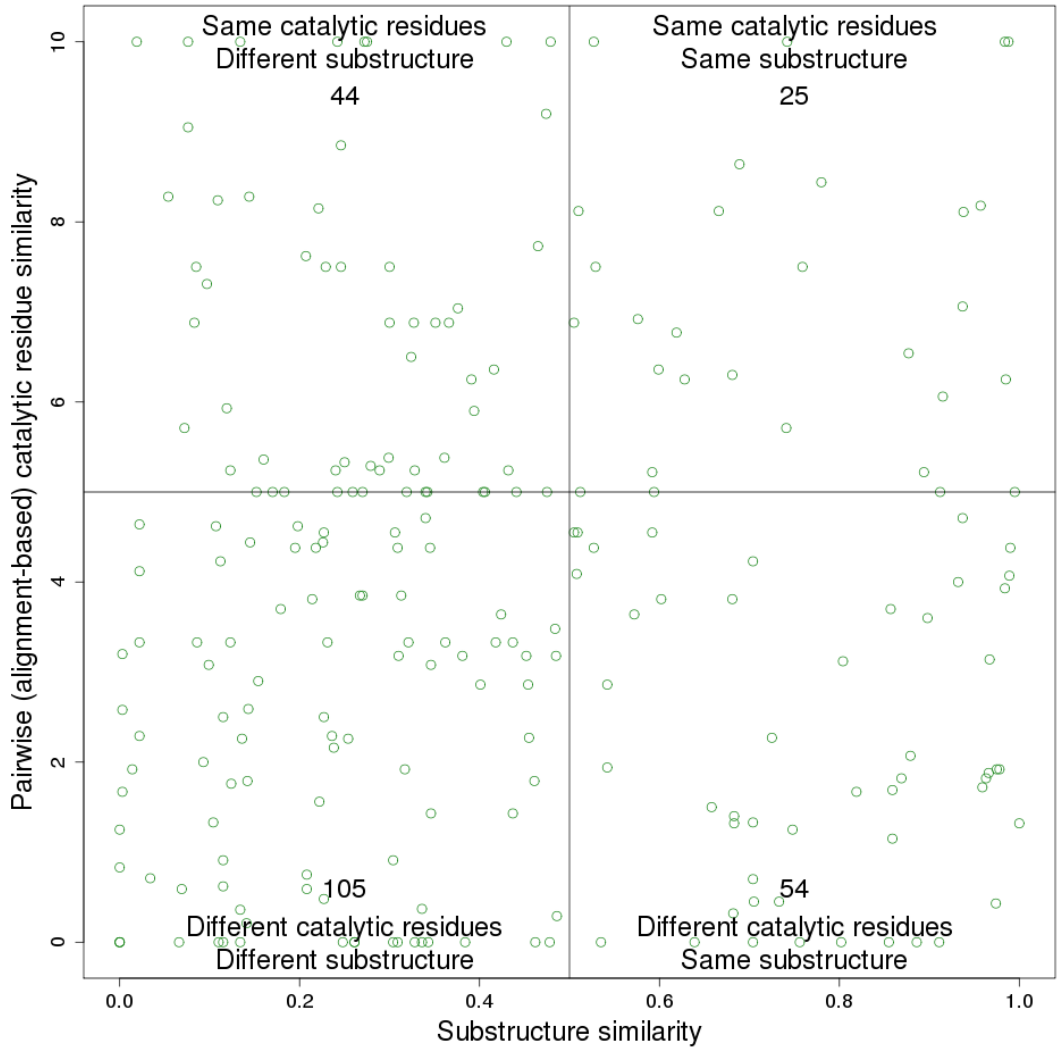
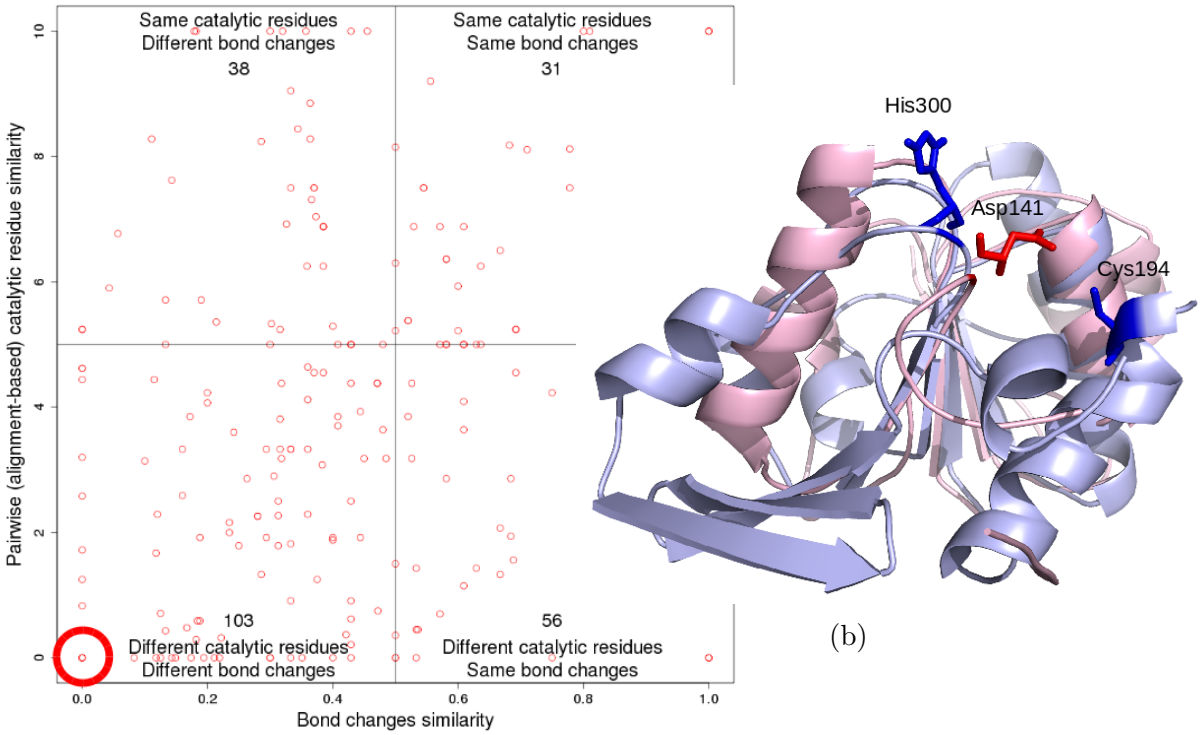


Figure 3.21: Correlating catalytic residue similarity with similarity in substructure.

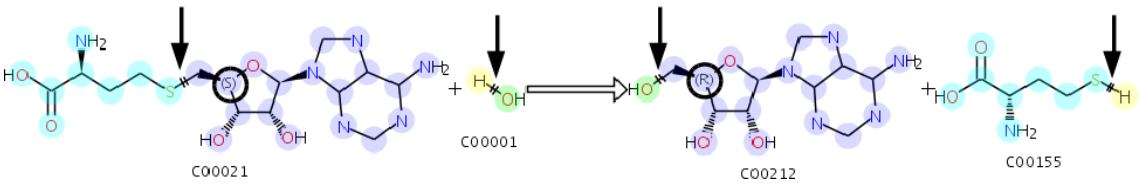
Reaction mechanism similarity data was not available for all functional family representative pairs. A total of 228 functional family representative domain pairs could be assigned a catalytic residue similarity score as well as a bond change similarity score. Since similarity in chemistry will be reflected by a similarity in the bond changes, Figure 3.19 is discussed in more detail below.

Different catalytic machinery, different reaction mechanism (bottom-left quadrant) In 103 functional family pairs (45.18%), changes in catalytic machineries are accompanied by differences in bond change. An example of two functional families within a superfamily which have diverged in both catalytic machinery and in reaction mechanism (i.e. bond change) is provided by the 'NAD(P)-binding Rossmann-like domain' CATH superfamily (ID: 3.40.50.720). Precorrin-2 dehydrogenase (P2DH) (EC 1.3.1.76) and adenosylhomocysteinase (AHS) (EC 3.3.1.1) both have catalytic residues present in loop regions: P2DH has a C-terminal Asp141 while AHS has a C-terminal His300 and a N-terminal Cys194 (see Figure 3.22 in red and blue, respectively).

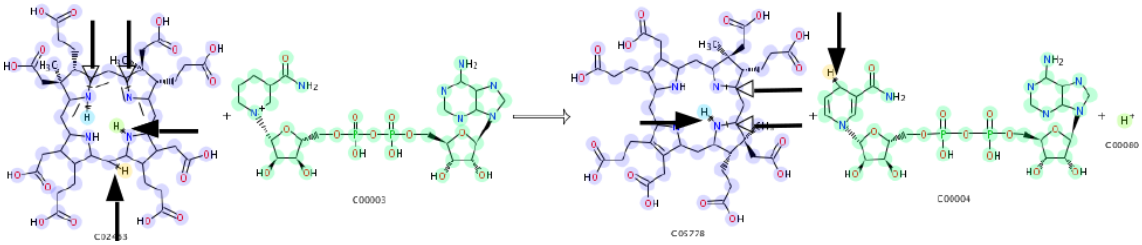


(a) (b)

Figure 3.22: Examining two functional family representatives from the 'NAD(P)-binding Rossmann-like domain' superfamily (ID: 3.40.50.720) which have no similarity in either their catalytic residues or in their reaction mechanisms, circled in Subfigure **(a)**. Subfigure **(b)** The superposition of the two functional family representative domains. They contribute to catalytic activity in adenosylhomocysteinase (CATH domain ID: 1b3rA01 in light blue) and precorrin-2 dehydrogenase (CATH domain ID: 1kyqA01 in light pink). Their catalytic residues are shown in blue and red, respectively.



(a) Adenosylhomocysteinase (EC 3.3.1.1, reaction ID: R00192): S-Adenosyl-L-homocysteine + H2O <=> Adenosine + L-Homocysteine



(b) Precorrin-2 dehydrogenase (EC 1.3.1.76, reaction ID: R03947): Precorrin 2 + NAD+ <=> Sirohydrochlorin + NADH + H+

Figure 3.23: Examining two functional family representatives from the 'NAD(P)-binding Rossmann-like domain' superfamily (ID: 3.40.50.720) which have no similarity in either their catalytic residues or in their reaction mechanisms. Subfigure **(a)** describes the chemical reaction catalysed by precorrin-2 dehydrogenase. Thick arrows identify the two bonds broken and the two bonds formed. Thick rings identify the single stereo change. Subfigure **(b)** describes the chemical reaction catalysed by adenosylhomocysteinase. Thick arrows identify the bonds cleaved/formed. The subfigures have been adapted from images taken from the EC-BLAST web site.

Different catalytic machinery, same reaction mechanism (bottom-right quadrant Surprisingly, and not reported on this scale in previous work, there are 56 pairs of functional families with different catalytic machineries associated with the same, or highly similar, bond changes. These are possible cases of functional convergence and we have examined some of the examples from this category below.

Within the CATH 'Vaccinia Virus protein VP39' superfamily, domains from the enzymes caffeate O-methyltransferase and catechol O-methyltransferase use different catalytic machineries but are associated with the same bond changes during the two different enzymatic reactions. Both enzymes have methytransferase activity (EC 2.1.1.-). The caffeate O-methyltransferase has a single catalytic histidine at position 269 while the catechol O-methyltransferase has a lysine residue at position 144 and a glutamic acid residue at position 199. The His269 and the Lys144 are at different sequence positions in the pairwise alignment of the functional family

representatives, but they are similar residue types and they closely co-locate in 3D space when superposed (see Figure 3.24). The CSA reports that His269 functions as a base to deprotonate an hydroxyl group on the reactant molecule (Furnham *et al.*, 2014) and it is likely that the Lys144 would have a similar basic role.

To examine the similarities between the two enzyme reactions further, the bonds cleaved and formed were identified through EC-BLAST. Figures 3.25a and 3.25b show the identical bond changes made.

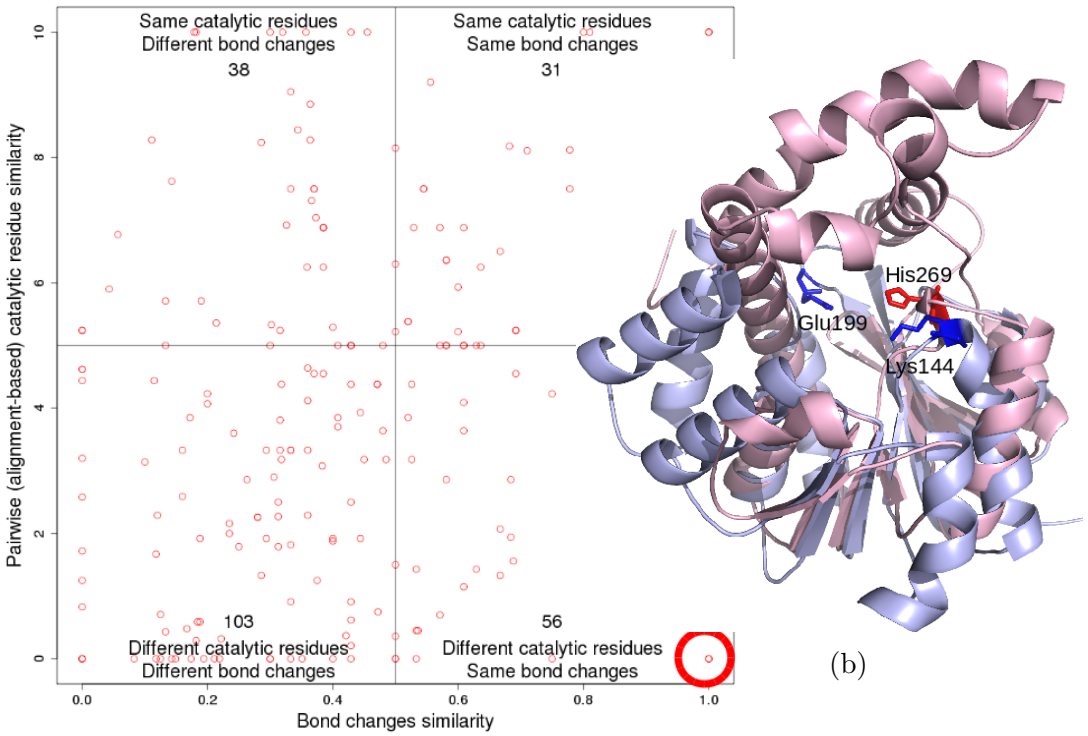
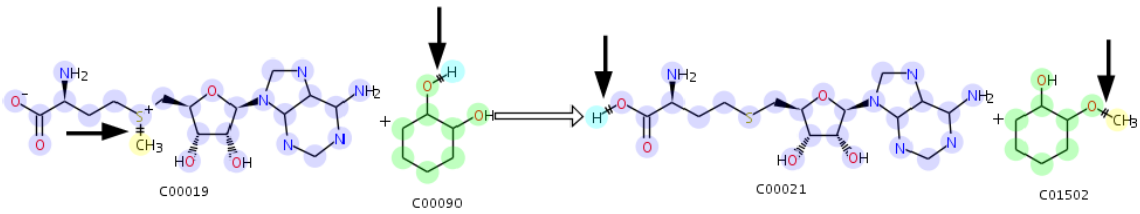
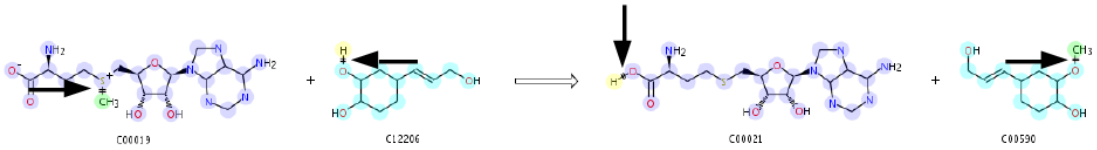


Figure 3.24: Examining two functional family representatives from the 'Vaccinia Virus protein VP39' CATH superfamily (ID: 3.40.50.150) which have no catalytic residue similarity but perform the same bond changes, circled in Subfigure **(a)**. Subfigure **(b)** shows the superposition of the two functional family representative domains. They are responsible for catalytic activity in caffeate O-methyltransferase (CATH domain ID: 1kywA02 in light pink) and catechol O-methyltransferase (CATH domain ID: 1vidA00 in light blue). Their catalytic residues are shown in red and blue, respectively.



(a) Catechol O-methyltransferase (EC 2.1.1.6, reaction ID: R00827): S-Adenosyl-L-methionine + Catechol <=> S-Adenosyl-L-homocysteine + o-Methoxyphenol



(b) Caffeate O-methyltransferase (EC 2.1.1.68, reaction ID: R06575): S-Adenosyl-L-methionine + Caffeyl alcohol <=> S-Adenosyl-L-homocysteine + Coniferyl alcohol

Figure 3.25: Examining two functional family representatives from the 'Vaccinia Virus protein VP39' CATH superfamily (ID: 3.40.50.150) which have no catalytic residue similarity but perform the same bond changes. Subfigure **(a)** describes the chemical reaction catalysed by catechol O-methyltransferase. Thick arrows identify the two bonds cleaved and the two bonds formed. Subfigure **(b)** describes the chemical reaction catalysed by caffeate O-methyltransferase. Thick arrows identify the same two bonds cleaved and the same two bonds formed. The subfigures have been adapted from images taken from the EC-BLAST web site.

Same catalytic machinery, different reaction mechanisms (top-left quadrant) The top-left quadrant of Figure 3.19 shows that a change in bond change reaction mechanism is sometimes achieved without a large, or any, change in the catalytic machinery. There are 38 pairs (16.7%) of homologous functional family representative domains that use similar, or identical, catalytic machineries to catalyse different enzymatic reactions.

Examining two enzymes with the same catalytic machinery in the Class I glutamine amidotransferase CATH superfamily (ID: 3.40.50.880), anthranilate synthase (AS) (EC 4.1.3.27) and carbamoyl-phosphate synthase (CPS) (EC 6.3.5.5) catalyse different chemical reactions with the same catalytic machinery. The partially-annotated approach found a common catalytic triad of Cys-His-Glu in both domains: i.e. there is a glutamic acid residue in the domain of CPS that was missing from the CSA. The presence of this Cys-His-Glu catalytic triad is supported by the work of Thoden *et al.* (1997), who found the small subunit of CPS (i.e. 1bxrB02) has

glutamine amidotransferase activity. This catalytic triad is found to be conserved in all domains with this enzymatic activity, e.g. the TrpG domain in AS (Spraggon *et al.*, 2001).

Figure 3.26 shows the identical positions of the catalytic residues between CATH domains IDs 1i7qB00 (AS) and 1bxrB02 (CPS), which represent subunits of these two enzymes. 1i7qB00 represents the TrpG subunit from AS; TrpG produces an intermediate in the formation of ammonia from a bound glutamine (Spraggon *et al.*, 2001). 1bxrB02 represents the small subunit from CPS that also produces ammonia by hydrolysing a bound glutamine. The ammonia is then used by the large subunit to produce carbamate (Thoden *et al.*, 1999).

Whilst both domains are annotated in the literature as catalysing two different EC enzymatic reactions the domains are actually performing the same function of converting glutamine to ammonia. Therefore, the fact that this functional family pair appear in the 'different reaction mechanisms' quadrant is the result of the EC number capturing the functional annotation at the whole-protein level rather than at the domain level.

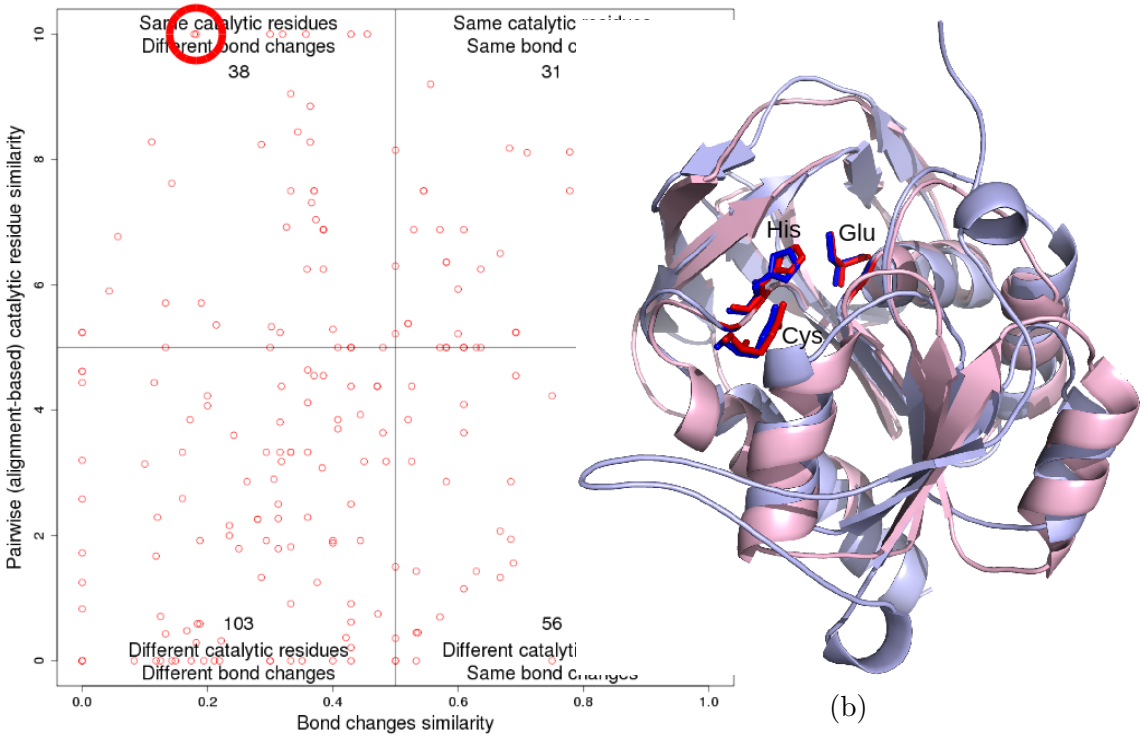
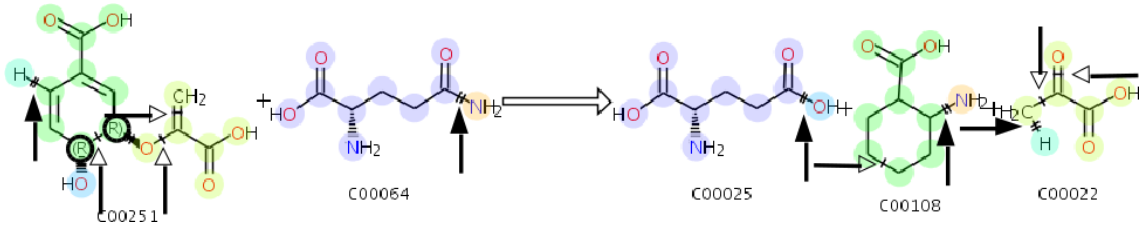
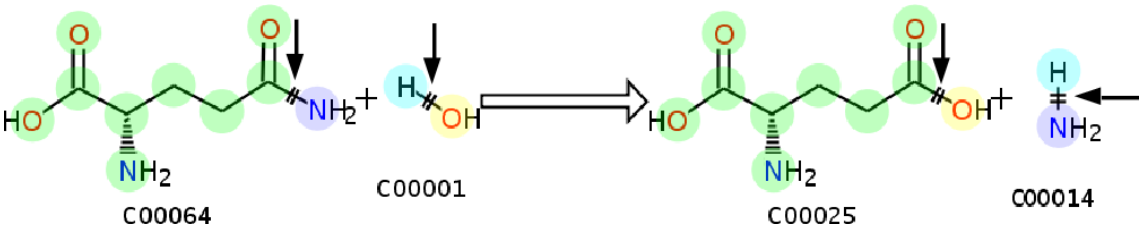


Figure 3.26: Examining two functional family representatives from the Class I glutamine amidotransferase CATH superfamily (ID: 3.40.50.880) which have a high catalytic residue similarity of ten but a low similarity in bond change (0.1790), circled in Subfigure **(a)**. Subfigure **(b)** shows the superposition of the two functional family representative domains. They are responsible for catalytic activity in anthranilate synthase (AS) (CATH domain ID: 1i7qB00 in light pink) from Serratia marcescens and carbamoyl-phosphate synthase (CPS) (glutamine-hydrolysing) from Escherichia coli (strain K12) (CATH domain ID: 1bxrB02 in light blue). Their catalytic residues are shown in red and blue, respectively.



(a) EC 4.1.3.27 (R00986): Chorismate + L-Glutamine <=>Anthranilate + Pyruvate + L-Glutamate



(b) EC 6.3.5.5 (R00256): L-Glutamine + H2O <=>L-Glutamate + Ammonia

Figure 3.27: Examining two functional family representatives from the Class I glutamine amidotransferase CATH superfamily (ID: 3.40.50.880) which have a high catalytic residue similarity of ten but a low similarity in bond change (0.1790). Subfigure **(a)** describes the chemical reaction catalysed by anthranilate synthase. Subfigure **(b)** describes the chemical reaction catalysed by carbamoyl-phosphate synthase. Thick filled arrows identify the bonds formed/cleaved. Thick white-headed arrows identify bond order changes. Thick black circles identify bond stereo changes. The subfigures have been adapted from images taken from the EC-BLAST web site.

Therefore, to avoid similar problems with other examples from this quadrant, we took a second example from a different superfamily, the Aldolase Class I superfamily, where the protein function is associated with a single domain. We selected the yeast L-lactase dehydrogenase (also known as flavocytochrome b2, or FCB) and spinach glycolate oxidase (GOX) enzymes. They use the same catalytic residues to perform different enzymatic reactions: FCB is a dehydrogenase (EC 1.1.2.3) whereas GOX is an oxidase (EC 1.1.3.15). Both are flavoprotein enzymes that catalyse the oxidation of different L-alpha-hydroxy acids. The flavoprotein enzyme family is well characterised with members sharing a number of properties. These enzymes bind the cofactor flavin and feature six conserved active site residues around this cofactor binding site in highly similar spatial positions (Macheroux *et al.*, 1993; Maeda-Yorita *et al.*, 1995). Figure 3.28 shows the superposition of the two domains and the six catalytic residues.

Interestingly, the first steps in the two enzymatic reactions (see Figure 3.29) are similar in that the lactate substrate of FCB and the glycolate substrate of GOX are both oxidised and the FMN cofactor is reduced. Subsequently the reactions diverge: in the FCB-catalysed reaction, the electrons from the FMN are transferred to a haem moiety (through a semiquinone intermediate) and then on to cytochrome c (Xia *et al.*, 1987). In the GOX-catalysed reaction, the reduced FMN is reoxidised by oxygen to hydrogen peroxide (Lindqvist, 1989).

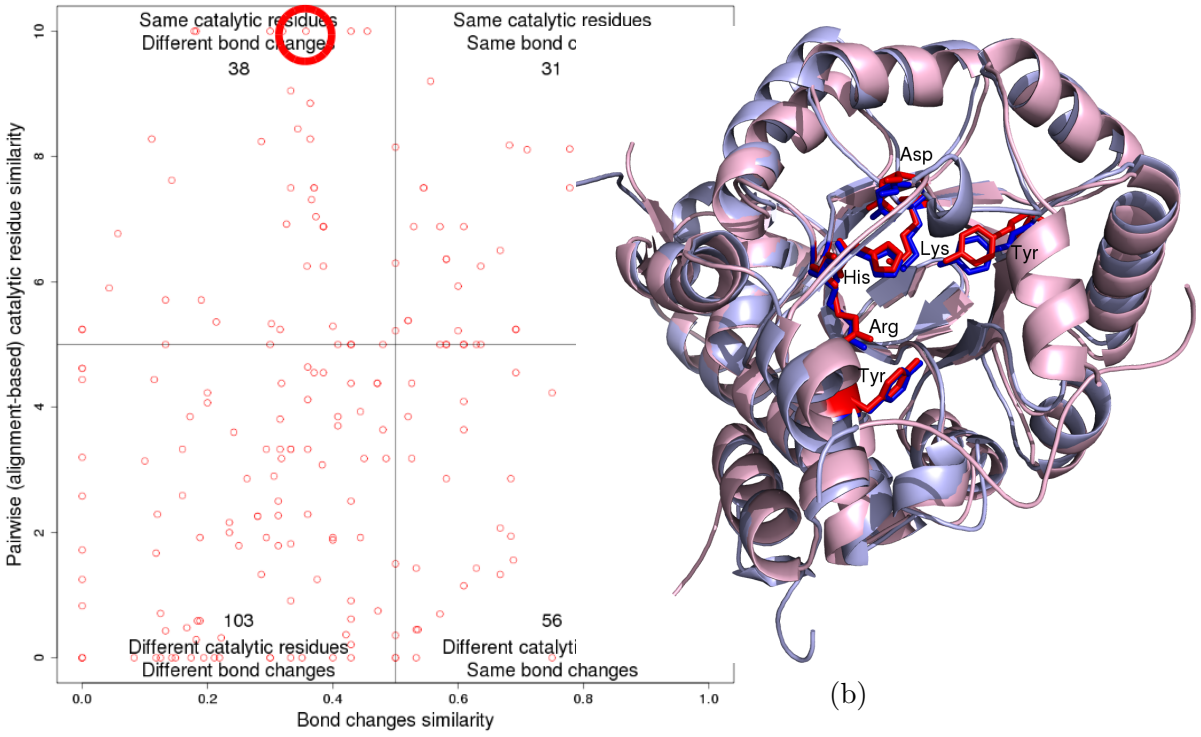
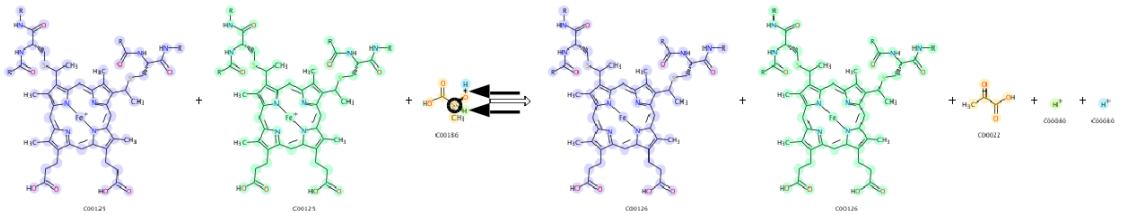
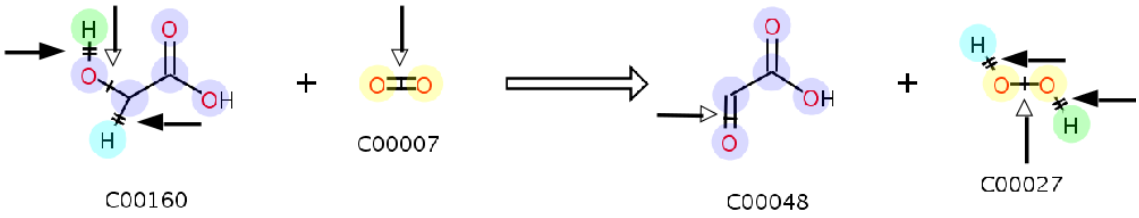


Figure 3.28: Examining two functional family representatives from the Aldolase Class I superfamily (ID: 3.20.20.70) which have a high catalytic residue similarity of ten and a low bond change similarity of 0.357, circled in Subfigure **(a)**. Subfigure **(b)** shows the superposition of the two functional family representative domains. They are responsible for catalytic activity in yeast flavocytochrome b2 (FCB) (EC 1.1.2.3, CATH domain ID: 1fcbA02 in light blue) and spinach glycolate oxidase (GOX) (EC 1.1.3.15, CATH domain ID: 1goxA00 in light pink). Their catalytic residues are shown in blue and red, respectively.



(a) EC 1.1.2.3 (R00196): (S)-Lactate + 2 Ferricytochrome c <=>Pyruvate + 2 Ferrocytochrome c + 2 H+

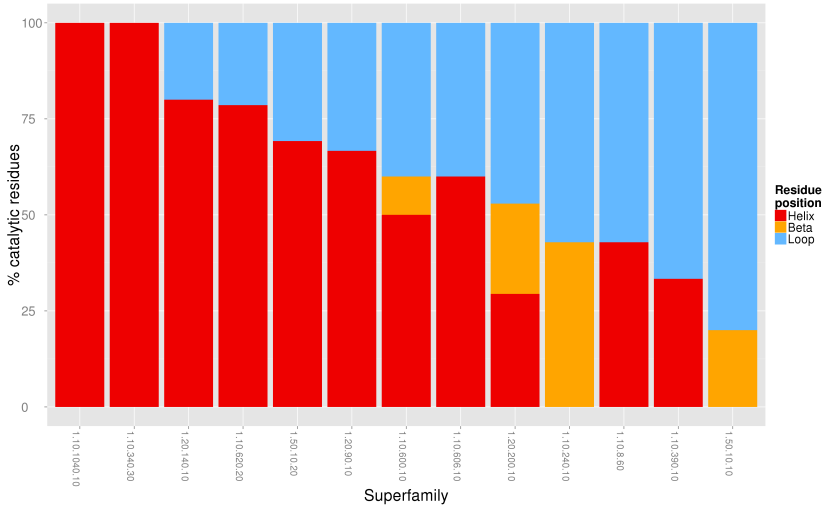


(b) EC 1.1.3.15 (R00475): Glycolate + Oxygen <=>Glyoxylate + Hydrogen peroxide.

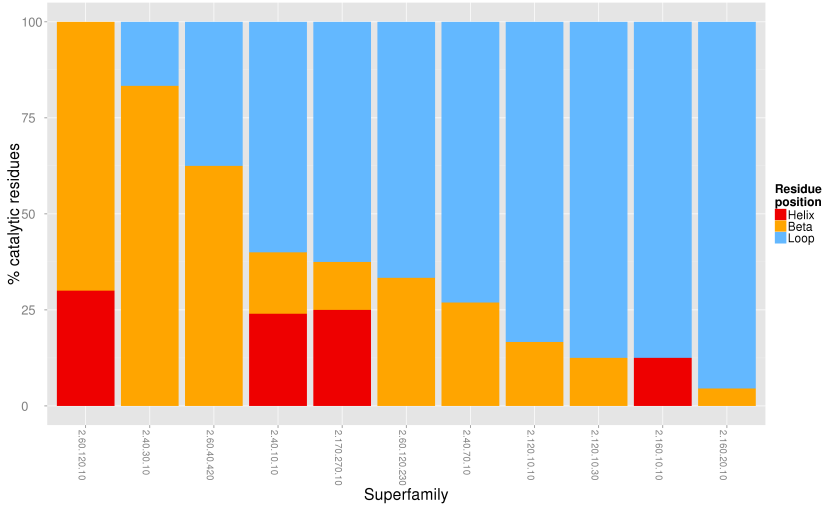
Figure 3.29: Examining two functional family representatives from the Aldolase Class I superfamily (ID: 3.20.20.70) which have a high catalytic residue similarity of ten and a low bond change similarity of 0.357. Subfigure **(a)** describes the chemical reaction catalysed by yeast flavocytochrome b2. Subfigure **(b)** describes the chemical reaction catalysed by spinach glycolate oxidase. Thick filled arrows identify the bonds formed/cleaved. Thick black circles identify bond stereo changes. Thick white-headed arrows identify bond order changes. The subfigures have been adapted from images taken from the EC-BLAST web site.

3.3.4 Examining whether catalytic residues are preferentially located in loop or secondary structure regions

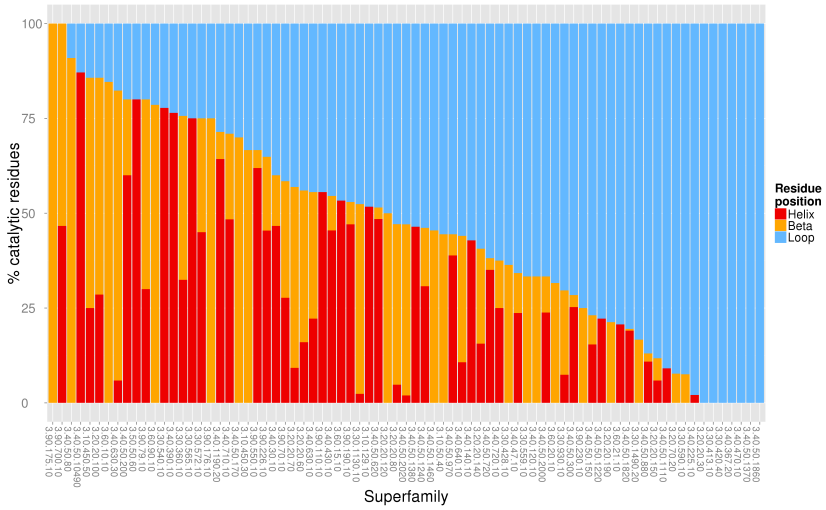
To examine whether catalytic residues are preferentially located in a particular structural element, the secondary structure properties of each catalytic residue in each superfamily were obtained through DSSP file information. Figure 3.30a shows that the majority of alpha-class superfamilies have $\geqslant 50\%$ of their catalytic residues in alpha-helices. On the other hand, the majority of beta-class and alpha-beta class superfamilies have $\geqslant 50\%$ of their catalytic residues in loop regions.



(a) Alpha CATH class.



(b) Beta CATH class.



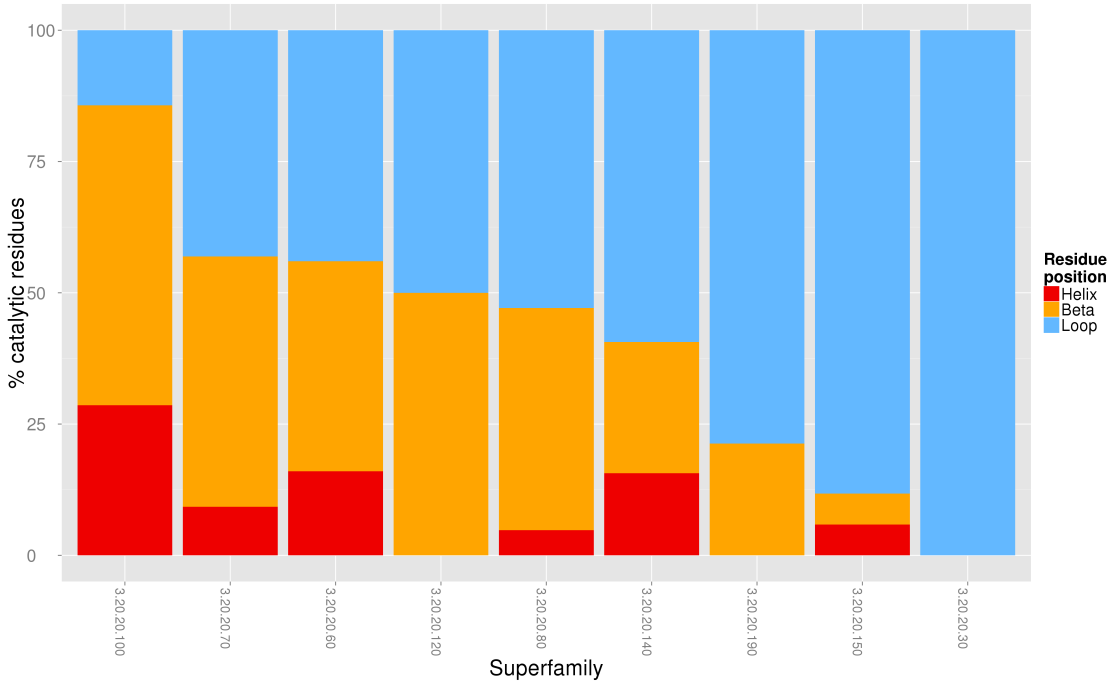
(c) Alpha-beta CATH class.

Figure 3.30: The location of catalytic residues within different types of secondary structure elements for superfamilies in different CATH class classifications.

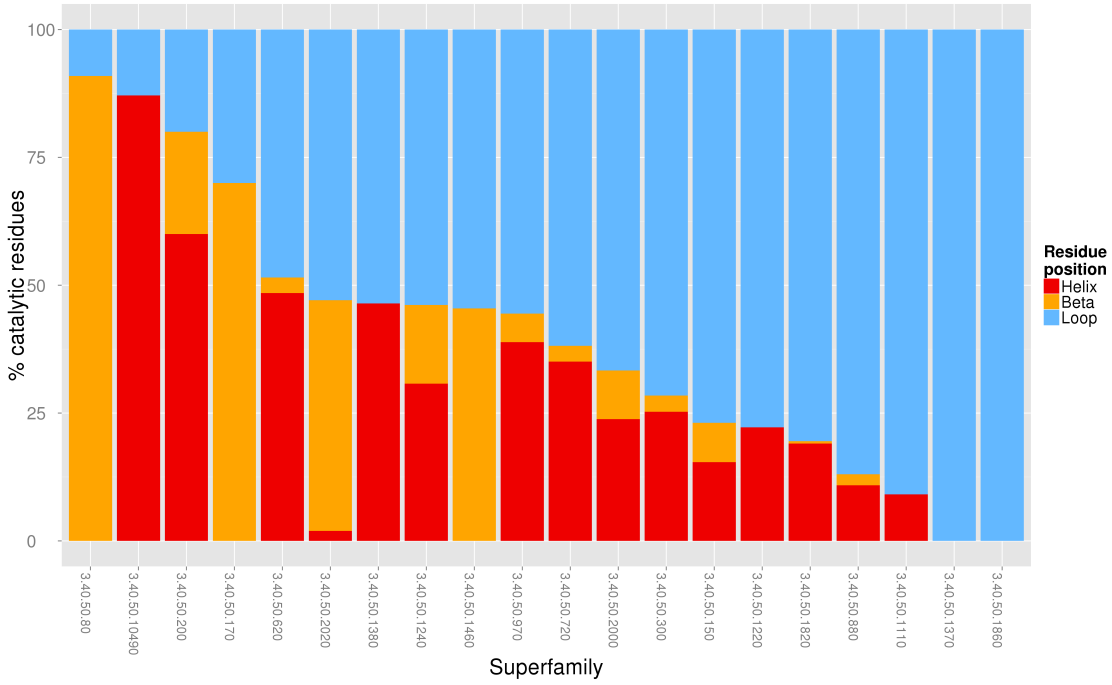
The Two Sample t-test statistical hypothesis test was used to examine whether there was a significant difference in the percentage of catalytic residues found in loop regions, based upon the CATH class or topology of the superfamily data. Beta and alpha-beta class superfamilies were found to have significantly higher percentage of catalytic residues in loops than the alpha class, than by chance ($t(22)=-2.4038$, $p < 0.02509$, and $t(88)=-2.3084$, $p < 0.02332$, respectively). No significant difference was found between beta and alpha-beta class superfamilies.

We then examined in closer detail the alpha-beta superfamilies with TIM barrel or Rossmann folds to investigate the preferential structural elements for catalytic residues in folds well known to support diverse functions. Almost 40% (29) of the 77 superfamilies with an alpha-beta CATH class classification belong to either the TIM barrel (CATH ID: 3.20.20) or Rossmann (CATH ID: 3.40.50) fold. The majority of these superfamilies have at least half of their catalytic residues in loop regions; 6 (out of 9, 67%) superfamilies with a TIM barrel fold have $\geqslant 50\%$ of their catalytic residues in loop regions (see Figure 3.31a) and 15 (out of 20, 75%) superfamilies with a Rossmann fold have $\geqslant 50\%$ of their catalytic residues in loop regions (see Figure 3.31b). These results agree with the findings from Bartlett *et al.* (2002) and Dellus-Gur *et al.* (2013), who found that catalytic residues tend to be found in the loop-regions of a protein structure, particularly if the protein has a highly stable structural scaffold and innovable fold, such as the TIM barrel fold.

However, the Two Sample t-test did not find the highly innovable TIM barrel and Rossmann folds to have significantly more catalytic residues in loops than by chance when compared to other folds in the alpha-beta class.



(a) TIM barrel fold (CATH Topology: 3.20.20).



(b) Rossmann fold (CATH Topology: 3.40.50).

Figure 3.31: The location of catalytic residues within different types of secondary structure for the TIM barrel and Rossmann fold groups.

3.4 Conclusions and future work

In this chapter changes in catalytic machineries have been explored between functional families in the same homologous superfamily. A total of 101 superfamilies were found to have CSA residue information in at least two functional families and were used in the analysis.

While the FunFam$_{SEQ}$ functional families were described in the previous chapter as containing structurally similar domains and being of good functional purity, further examination has highlighted that some functional families contained fairly structurally diverse domains and multiple EC terms. The FunFam$_{SEQ}$s were therefore split into FineFams, where each FineFam has a single EC number annotation. Following this splitting process, greater similarity in catalytic machinery was observed within the functional families.

When comparing the catalytic machinery of two functional family representatives, an RMSD cut-off of 5 Å was used to ensure that the two superposed domains were not structurally diverse, as this would affect the sequence alignment. A total of 79 CATH enzyme domain superfamilies had functional family representatives that could be aligned well to compare their catalytic residues, and these were compared with regards to their physicochemical properties. Two different approaches were used, which either scored pairs of catalytic residues that were both annotated by the CSA (i.e. the 'fully-annotated' approach) or that scored pairs where at least one residue in the pair was annotated by the CSA (i.e. the 'partially-annotated' approach). The latter approach was used to identify potential catalytic residues that are missing from the CSA. Considerable variation in catalytic machineries has been observed across the majority ($\geqslant 60\%$) of superfamilies.

We examined whether a change in catalytic machinery is accompanied by a change in reaction mechanism. Information on bond change, reaction centre, and substructure similarity was used to characterise reaction mechanism similarity. No clear correlation was found. Comparing functional families across a superfamily, the smallest percentage of comparisons showed the same catalytic residues and the

same reaction mechanism, whether measured by bond change, reaction centre, or substructure similarity. Perhaps unsurprisingly, the largest percentage showed differences in catalytic residues accompanied by differences in chemistry. Interestingly, quite a large proportion of functional family comparisons revealed that the same catalytic machineries were supporting different chemistries. Most of the time, changes are occurring in the substrate, this suggests that the chemistry enabled by a particular arrangement of catalytic residues has been exploited to operate on different substrates or in different enzyme complexes/metabolic pathways. It would be interesting to examine these in more detail in the future.

Perhaps most interesting was the observation that in nearly one quarter of functional families, we observed that different catalytic machineries were performing the same bond changes. For example, we reported two domains in caffeate O-methyltransferase and catechol O-methyltransferase which have no similarity in their catalytic machineries, the former uses a single histidine and the latter uses a glutamic acid and a lysine, yet each cleave/form one O-H hydrogen bond and cleave/form one sulphur-carbon bond. This observation is on a much larger scale than reported by previous studies and supports early studies of Babbitt (Babbitt and Gerlt, 1997) and Thornton (Todd *et al.*, 2001) which reported conservation of one or more steps in the reaction pathway. A large proportion of these functional family pairs were associated with changes in substrate and therefore may reflect changes in residues that are associated with substrate interactions and stabilising transition states. In other words, the key catalytic residues may be similar in these functional family pairs but changes of residues (marked as catalytic residues by the CSA) may be necessary for specific chemical interactions with the different substrates during the reaction steps. Even more surprising and interesting, was the observation of quite a significant proportion of pairs of functional families with different catalytic machineries and the same chemical reaction performed on the same substrate. In this scenario, it is possible that mutations in residues in the active site are subject to neutral drift, which can remove catalytic properties, and are then followed by

further mutations that restore catalytic activity by providing a new arrangement of catalytic residues which can also perform the same reaction.

To compare the reaction mechanisms of two functional family representatives, they must have an EC number as this will be associated with at least one IUBMB enzyme reaction code. EC-BLAST uses the first IUBMB reaction code to calculate the reaction mechanism similarity. It has not been possible to compare all pairs of chemical reactions as EC-BLAST requires that a chemical reaction is chemically balanced and this was not possible with a number of entries.

We also examined whether catalytic residues prefer being located within a particular structural element. We used the full set 101 CATH superfamilies that have two or more functional families with CSA residue information. The majority of superfamilies in each CATH class had at least half of their catalytic residues in loop regions with 61.54%, 72.73%, and 51.95% of superfamilies in the alpha, beta, and alpha-beta classes, respectively. The TIM and Rossmann superfamilies make up 40% of the 77 analysed alpha-beta superfamilies and the majority (88.76%) have at least half of their catalytic residues in loop regions.

Beta and alpha-beta class superfamilies are perhaps more likely to have catalytic residues in their loop regions, as they are thought to possess stable structural scaffolds comprised of beta sandwiches and beta barrels. These scaffolds often support many different functions and all feature loops clustered around the active site. In the alpha-beta superfamilies, the active site is generally found at the C-terminal ends of beta strands, as well as in the loops that link the beta strands with alpha helices. These loops do not contribute to protein stability (Dellus-Gur *et al.*, 2013) and therefore insertion of residues at these positions can easily change the geometric and physicochemical properties of the active site. We are currently analysing the preference for catalytic residues in loop regions in all superfamilies with structural scaffolds comprising beta sandwiches or beta barrels.

Chapter 4

Functional Analysis of the Human Oral Metagenome

4.1 Introduction

4.1.1 Metagenomics and the microbiome

Metagenomics is a relatively new field that analyses DNA extracted from any given environment to examine the species composition and functional ability of the microbial community found in that environment. The term 'metagenomics' was first used by Handelsman *et al.* (1998). The first large-scale metagenomic studies were carried out in 2004 by Tyson *et al.* (2004) and Venter *et al.* (2004) who studied the genomes of microorganisms within an acid mine drainage biofilm, and within ocean water samples, respectively.

Before metagenomics, genome sequences could only be studied if the organism could be cultivated in the laboratory. This provided a biased view of bacterial life on earth as it is estimated that only approximately 1% of all microbes can be cultured (Hugenholtz *et al.*, 1998; Foerstner *et al.*, 2006; Kristiansson *et al.*, 2009). As the species within the sampled environment do not require cultivating in the laboratory, the new sequencing technologies exploited by metagenome studies can provide a relatively unbiased view of the community structure, also termed the microbiome, and its functional repertoire.

Following the large-scale studies from Tyson *et al.* (2004) and Venter *et al.* (2004), whole-genome shotgun sequencing and sequence read assembly has been increasingly used to sequence and study mixed microbial communities.

4.1.2 Experimental characterisation of the metagenome

Characterising microbial species Before the 1980s, researchers would characterise a species by publishing summary tables of phenotypic traits, for example whether the bacterium was gram-positive or gram-negative (Woese, 1987; Clarridge, 2004; Kim *et al.*, 2011). There was frequently no match to the query species and an educated guess had to be made. Such methods were dependent on being able to culture the organism in the laboratory, and as more and more organisms were being discovered, this method proved subjective and unreliable (Clarridge, 2004). As previously mentioned, there are also few well-characterised microbes.

In 1965, Dubnau *et al.* discovered that the 16S rRNA gene in Bacillus species was highly conserved. Pioneering work from Woese (1987) then showed that the 16S rRNA gene sequence could be used across all kingdoms to identify species through sequence similarity searches. This was due to the 16S rRNA gene being one of the few ubiquitous and highly evolutionarily conserved genes. The 16S rRNA gene is a common housekeeping genetic marker approximately 1,550 bp in length and contains variable and conserved regions. To create a 16S rRNA profile, amplification primers are typically used to target three of the nine hypervariable regions in the 16S rRNA gene: V2, V3, and V6 (Chakravorty *et al.*, 2007).

The Pace group used this gene to construct the first phylogenetic analysis of a microbial community, a marine picoplankton community (Schmidt *et al.*, 1991). This led to the development of widespread taxonomic classification through 16S rRNA gene sequence analysis (Clarridge, 2004).

Characterising microbial protein function When studying genes of functional interest, there are two types of methodologies traditionally used: function-based and sequence-based metagenomics. Function-based metagenomics involves cloning environmental DNA into expression vectors and propagating them into a host. An activity screen is used to filter for a function of interest. The active clones that perform the function of interest are sequenced and the genes and their protein products

analysed (Chistoserdova, 2010). This is useful when looking for proteins from new, or poorly characterised, species with a particular function.

4.1.3 Next-generation sequencing

DNA sequencing characterises the sequential order of different nucleotide bases in a length of DNA. The first techniques to achieve this were based upon the chemical modifications of DNA, developed by Maxam and Gilbert (1977) and the chain-termination method developed by Sanger *et al.* (1977). The latter method, termed 'Sanger sequencing', was the first to be automated, which allowed researchers to sequence large quantities of DNA faster and more cheaply. The technology was developed over the next two decades and became the first method to sequence the full human genome in 2001 (Lander *et al.*, 2001). Sanger sequencing is similar to natural DNA replication as it uses DNA polymerase to elongate complementary strands of short primers. Different labels are used to identify the four different dideoxynucleotides so that an addition of one can be detected through a detectable chain termination (Mutz *et al.*, 2013). This method is associated with a low sequencing error rate of $\sim 2\%$ and read lengths of up to ~ 2000 bp (Nagarajan and Pop, 2013).

Pyrosequencing, also known as sequencing by synthesis, is a DNA sequencing technique developed in the late 1990s by Ronaghi *et al.* (1996). This technique brought an attractive alternative to Sanger sequencing through its ability to perform real-time sequencing that was simple, automated and faster than previous methods. As DNA polymerase moves along an immobilised single stranded template of DNA, the four different nucleotides are sequentially added in solution, and if incorporated a flash of light is detected. The pyrophosphate released from the DNA polymerase-catalysed reaction forms ATP, which is then used in the ATP-dependent conversion of luciferin to oxyluciferin. The production of oxyluciferin causes a pulse of light, whose amplitude is directly related to the number of nucleotides incorporated (Ronaghi *et al.*, 1996; Petrosino *et al.*, 2009). DNA pyrosequencing is generally only able to sequence DNA fragments up to 100-200 bases whereas from 2005, 454 Life Sci-

ences released a high-throughput pyrosequencing technique, which can now sequence fragments up to approximately 750 bases (Margulies *et al.*, 2005; Glenn, 2011). This technique is known as 454 sequencing and it was the first next generation sequencing method. While longer reads are produced with fast run times, the reagent costs are high and there are high error rates associated with homopolymer repeats and duplicate reads (Glenn, 2011). Nagarajan and Pop (2013) reported a sequence error rate of $\sim 4\%$ for this technology.

Illumina next generation sequencing (NGS) (initially developed in 2007 by Solexa) also uses a sequencing-by-synthesis method and was the first next generation short-read sequencer (Bentley *et al.*, 2008). The DNA of an amplified library of fragments is sequenced using reversible dye terminators. In this method, all four nucleotides can be added at the same time in each cycle as each carries a different fluorescent label. A nucleotide is added by DNA polymerase, then the unincorporated nucleotides are washed away and an image is taken to identify the fluorescent signal. The fluorescent group is then cleaved and the 3'-hydroxyl group is chemically de-blocked so that the next nucleotide can be incorporated. Up to 150 nucleotides can be added in this way. Most Illumina reads are reported to have an error rate of 0.5%, i.e. 1 error in 200 bases) (Mardis, 2013). Nagarajan and Pop (2013) also reported a low sequencing error rate of $< 2\%$. These errors can be a result of phasing, which is where the de-blocking process is incomplete, or where a blocking group is missing. Errors can also result from fluorescence interference noise, which can occur when a fluorescent group has not been cleaved from a previous cycle (Mardis, 2013).

Most recently, instruments have been developed that are able to sequence individual strands of metagenomic DNA in real time. Pacific Biosciences, known as PacBio, was developed in 2009 (Eid *et al.*, 2009) and made commercially available in 2010. Starlight is another single-molecule sequencing technique however it is still under development and is not commercially available. In PacBio, each nucleotide has a different fluorescent label that is detected as soon as it is cleaved during synthesis (Glenn, 2011). While the reads produced are 964 bases on average, it has the

highest error rates compared to other NGS techniques of $\sim$ 18% (Metzker, 2010; Nagarajan and Pop, 2013).

A number of metagenomic studies have used 454 pyrosequencing since its release in 2005. A single run of 454 pyrosequencing allowed for the analysis of a 13 Mb sequence of 28,000-year-old mammoth in 2006 (Poinar *et al.*, 2006). Since then, projects using 454 technology have investigated the metagenomes of soils (Leininger *et al.*, 2006), a coral holobiont (Wegley *et al.*, 2007), and nine biomes (Dinsdale *et al.*, 2008), which include stromatolites, fish gut, fish ponds, mosquito viriome, chicken gut, bovine gut and marine viriome (Hugenholtz and Tyson, 2008). Recent years have seen a focus on the sequencing of the human microbiome using next-generation technologies, with projects such as the HMP as mentioned in Section 4.1.9. Another recent example used the Illumina sequence reads to establish a human gut microbial gene catalogue (Qin *et al.*, 2010).

As there are error rates associated with all methods of DNA sequencing a method was developed to calculate the reliability of each base-call through a quality score. The program Phred was developed to estimate the probability of error for each base-call (Ewing and Green, 1998). Log-transformed error probabilities are used to calculate a quality value (q) (see Equation 4.1).

$$q = -10 \text{ x } log_{10}(p) \tag{4.1}$$

p represents the estimated error probability for a given base-call. A high quality value corresponds to a low error probability. For example, a Phred quality score of 30 corresponds with a 1 in 1000 chance of an incorrect base call (Ewing and Green, 1998).

4.1.4 Computational methods to classifying microbiome species

The species present within an environment's bacterial community can be identified using two different types of sequence data: 16S rRNA gene sequences or whole-genome shotgun sequences.

In metagenomics, a whole-genome shotgun sequencing approach is taken to sequence whole genomes rather than a single gene. These genomic sequence reads are typically scanned against a database containing 16S rRNA gene data, such as the Ribosomal Database Project (RDP) to classify microbial species, though they can also be scanned against whole protein sequences in resources including UniProtKB, GenBank and RefSeq.

From these approaches, microbial species can be identified. However if there are any novel microbial species in a metagenome data set, they cannot be classified into the species taxonomy. In these cases, sequences can be clustered into groups of operational taxonomic units (OTUs) based upon their sequence similarities (Wu *et al.*, 2013). While this is advantageous for examining biodiversity, there are a number of disadvantages to this approach. There is no phylogeny hierarchy, different algorithms cluster OTUs in different ways, and very short sequences will rarely overlap with sequences in the database being searched to calculate sequence similarity (Wu *et al.*, 2013).

4.1.5 Computational methods to predict protein function

As discussed in the previous work chapters, protein function can be predicted in a number of ways. With metagenome data, we are presented with a large volume of DNA sequences which are typically annotated using protein sequence homology approaches.

There are three databases which are frequently mentioned in the literature when assigning functional annotations to metagenome sequences. KEGG Orthology (KO) terms from KEGG, which are manually defined sets of orthologous sequences for all proteins and functional RNAs that each represent a node in a KEGG pathway (Kanehisa *et al.*, 2014). The Cluster of Orthologous Groups (COGs) database (Tatusov *et al.*, 2003), contains clusters of prokaryotic and eukaryotic orthologues. The evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOGs) database contains orthologous groups clustered using sequences

from complete high-quality genome projects. Sequences are clustered using whole-protein sequence similarity information and orthologous proteins are identified within the clusters. Functional annotations are then retrieved for these orthologous proteins using any associated GO terms, membership to KEGG pathways, and the presence of domains from SMART and Pfam (Jensen *et al.*, 2008; Powell *et al.*, 2014).

CATH superfamilies and functional families, and members of InterPro such as Pfam, are additional sources of functional annotation, which have been discussed in more detail in previous chapters.

4.1.6 Web servers providing metagenome analysis protocols

A number of web servers have relatively recently become available which promote the public availability of metagenomic data and the standardisation of the analysis workflow. Data can be uploaded to a web server which processes and analyses the data to predict species in the microbiome and to provide protein function annotations. These resources include: MG-RAST (Metagenome rapid annotation using Subsystems technology) (Meyer *et al.*, 2008), Galaxy (Kosakovsky Pond *et al.*, 2009), IMG/M (Integrated microbial genomes and metagenomes) (Markowitz *et al.*, 2012), and EBI Metagenomics (Hunter *et al.*, 2014).

Following the submission of metagenome sequence data, all four web servers first perform quality analysis. Sequences can be: clipped to remove bases with low quality scores at sequence ends, removed if they contain too many ambiguous bases (i.e. 'N's), removed if they are duplicate reads, and removed if they are too short or too long.

The Galaxy web server has the most flexibility of the four in how the analysis is performed as the user can pick and chose each step of the analysis from a wide range of options. However there is also a metagenomics workflow which has been set up that takes the user through the steps to select high quality sequence reads and classify microbiome species (Kosakovsky Pond *et al.*, 2009). The other three web servers pre-define how the sequence data will be filtered and analysed.

The MG-RAST, EBI Metagenomics, and IMG/M pipelines predict genes in the filtered sequence reads using FragGeneScan (Rho *et al.*, 2010). IMG/M however also uses additional methods to predict different classes of genes, including non-coding RNA genes and protein-coding genes. Non-coding RNA genes here include tRNA genes and rRNA genes. Aside from FragGeneScan, protein-coding genes are predicted using *ab initio* programs: GeneMark v2.6, MetaGene, and Prodigal (Markowitz *et al.*, 2012).

Different methods are used by the web servers to classify the microbial species. EBI Metagenomics classifies species in the microbiome by scanning 16S rRNA sequences against the SILVA (Pruesse *et al.*, 2007) and Greengenes (DeSantis *et al.*, 2006) 16S rRNA sequence databases in QIIME v1.5 (Quantitative insights into microbial ecology) (Caporaso *et al.*, 2010). IMG/M classifies the species and creates a phylogenetic distribution based upon the results of BLASTing the predicted protein-coding genes against NCBI COGs. MG-RAST on the other hand scans sequences against data from the SILVA, Greengenes, and RDP (Wang *et al.*, 2007) databases, which have been integrated into a single non-redundant database called 'M5rna' (Meyer *et al.*, 2008).

The different metagenome pipelines use different combinations of functional resources, described earlier in Section 4.1.5, to functionally annotate the sequences. In the EBI Metagenomics pipeline, the protein-coding regions are scanned against InterProScan to functionally annotate the sequences. Within the InterPro resources, Gene3D, Pfam, TIGRFAM, PRINTS, and PROSITE patterns are currently used (Hunter *et al.*, 2014). In the IMG/M, genes are scanned against Pfam, TIGRFAM, and NCBI COGs. KEGG Orthology (KO) terms and EC numbers are also assigned through similarity searches (Markowitz *et al.*, 2012). In MG-RAST, sequences are scanned against their non-redundant database, 'M5nr', which incorporates sequences from GenBank, SEED, IMG, UniProt, KEGG, and eggNOGs databases. Sequences are assigned to different functional hierarchies using the SEED Subsystems, IMG terms, COGs, eggNOGs, and ontologies such as KEGG Orthology terms and GO

terms (Meyer *et al.*, 2008).

4.1.7 Sequence read assembly

As the sequenced reads of DNA are generally short, it can be difficult to reliably predict species within the microbiome and thereby, also the protein function. As most of the bacteria in the human gut cannot be cultured and therefore do not have reference genomes, *de novo* sequence assembly is used to assemble to sequence reads into longer contiguous sequences, or contigs, that represent the genomic content of bacteria in the metagenome.

Sequence assembly is challenging due to a number of factors. First, sequence reads that share the same repeating sequence can be impossible to distinguish between. Second, sequence read error is another factor; the algorithm needs to tolerate imperfect sequences in order to align and assemble sequences, however it must not be too lenient otherwise an assembly could be formed from sequences originating from different species, i.e. forming a chimeric sequence assembly. Third, the coverage of all the species in the metagenome can vary due to biases in sequencing technology, where low coverage can lead to gaps in assemblies.

An assembly is reported by the size of the contigs produced. Size statistics include the maximum length, average length, and combined total length of the contigs. The N50 statistic is also computed, which represents the length of the smallest contig in a collection of contigs that contain at least 50% of the assembled bases (Miller *et al.*, 2010). These measures can be used to gauge how well the assembly protocol performed, however it is very difficult to assess the exact accuracy unless there are reference genomes to compare against.

4.1.8 Complications with metagenome analysis

There are a number of complications involved with analysing metagenome sequence data. For example, each member of the microbiome is present at varying levels of abundance and their respective genome sizes determine how deep the sequence

coverage is. Therefore sequence reads from uncommon species are under-represented and it is difficult to obtain information from these species (Gill *et al.*, 2006).

Another issue is that the microbial genes, rather than expressed proteins, are being studied. This means that the genetic material might not be expressed by the microbial community, or it may be expressed differently due to post-translational modifications. The introduction of metatranscriptomics is one way of overcoming these issues by sequencing the mRNA transcripts produced by the members of the microbial community. The sequencing of transcripts is performed by next generation sequencing technology, which has been termed 'RNA-Seq'.

The comparison of metagenome data also comes with its own complications, discussed in detail by Raes *et al.* (2007). They suggest a number of factors to watch out for, which are based upon the biology of the environment being studied and also upon the technology being used. Related to the former set of factors, the analysis can be affected by the number of different species in each sample being compared. For example, if one sample contains less than ten species and a second sample contains hundreds of species, the species in the first sample will have much higher coverage of sequence reads. This would lead to better species and function prediction. Technical issues to be aware of include the use of different DNA extraction protocols, which may filter out certain organisms before the analysis can be carried out (Raes *et al.*, 2007).

4.1.9 The human metagenome

It has been estimated that the microorganisms that live inside and on the human body are at least ten times greater in number than all of the human somatic and germ cells in a single body (Turnbaugh and Gordon, 2008). This total number of human and microbial cells is estimated to be more than 10^{14} cells (Darveau, 2010). These microbial symbionts are extremely advantageous to humans as they provide traits that humans no longer need to maintain (Gill *et al.*, 2006). Examples of these traits include functional contributions to the human gut such as the harvesting of

otherwise inaccessible nutrients and/or sources of energy from the host diet, synthesis of vitamins, metabolism of xenobiotics, the renewal of gut epithelial cells and the development and activity of the immune system. Cardiac size is also affected by the microbiome (Turnbaugh and Gordon, 2008).

The human microbiome has been the subject of many studies. Perhaps the largest study carried out was by the Human Microbiome Project (HMP), an international collaboration between groups in the United States, Europe and Asia, which ran from 2007 to 2014 (NIH HMP Working Group *et al.*, 2009; Turnbaugh *et al.*, 2007). Samples were taken from all over the body, including the: blood, eye, gastrointestinal (GI) tract, oral cavity, skin, airways, and urogenital area. The HMP was an experimental extension of the Human Genome Project with three major goals: 1) to use high-throughput technologies to characterise the human microbiome through the sampling of multiple body sites from at least 250 healthy volunteers, 2) to discover any associations between microbiome perturbations and human health, and 3) to provide a standardised data resource and new technological approaches for worldwide and large-scale application (NIH HMP Working Group *et al.*, 2009). The global aim was to investigate and demonstrate how human health could be improved through the study of the human microbiome.

The human gut metagenome Another human metagenome-based consortium was the MetaHIT (Metagenomes of the human intestinal tract) project, which ran from 2008 to 2012, that aimed to understand the link between the gut microbiome and chronic human diseases. Qin *et al.* (2010) reported the first large scale study in association with the MetaHIT project. They extracted the DNA from 124 healthy, overweight, and obese European adults, as well as inflammatory bowel disease (IBD) patients from Denmark and Spain. A catalogue of all the microbial genes was published, describing between 1,000-1,150 prevalent bacterial species within the entire cohort. Each individual was reported to contain at least 160 prevalent bacterial species. The second major study done in collaboration with MetaHIT and the HMP by Arumugam *et al.* (2011) found that bacterial species from 33 European samples

formed three clusters, termed enterotypes, where each was predominant in either the Bacteroides genera, Prevotella, or Ruminococcus. They repeated the analysis with large-scale study samples from US and Europe and again identified three distinct clusters, where the third cluster was prevalent in Clostridia (Arumugam *et al.*, 2011). Following such studies, the human gut has been described as the most complex human metagenome with the highest number of bacterial species (Dave *et al.*, 2012).

The human oral metagenome The presence of bacteria in the human mouth was first reported by Antonie van Leeuwenhoek (1632-1723), who studied the "white stuff" between his teeth using his home-made microscopes (Porter, 1976).

The mouth is a warm and moist environment that provides many different types of surfaces available for bacterial colonisation. It has been shown that bacteria are found shortly after birth and are thought to be acquired from the child's parents/carers as very few of the bacteria are free-living (Gibbons and Houte, 1975). A large change in the bacterial composition is observed when the first teeth emerge, as more anaerobic bacteria such as Fusobacterium and Bacteroides species are observed (Gibbons and Houte, 1975; Wilson, 2004). Recent studies have suggested that the oral microbiome is the second most diverse in terms of species, after the gut microbiome (Wade, 2013).

While many surfaces are kept relatively bacteria-free due to the cleansing action of swallowing, there are a number of protected surfaces and crevices: the occlusal surfaces of teeth (i.e. the surface tip of a tooth that comes into contact with a tooth from the other jaw), between the teeth, and along the gingival margin (i.e. gum line). While many bacteria are observed on the teeth, there were thought to be less around the mucosal glands due to the shedding of epithelial cells (Gibbons and Houte, 1975). It was concluded that the amount, and type, of secretions in the mouth played a major role in the removal of bacteria, food particles, and other substances, together with forces from tongue movements and mastication (Gibbons and Houte, 1975).

Those bacteria that manage to adhere form thick biofilms, which are aggregates of microorganisms that use specific adhesins and receptors to interact with the host organism and numerous other microbes (Foster *et al.*, 2003; Shestakov, 2011). As the mouth is a hostile environment to microbes (due to rapid changes in temperature, pH, and clearance from swallowing) only a small proportion of those entering the oral cavity are able to attach and survive. Bacteria have also evolved to cope with the innate and adaptive responses of the host defence system. Biofilms allow the microbes to remain attached to the oral surface and they prevent bacterial removal during swallowing. Biofilm composition and metabolism has been observed to change according to temperature, pH, nutrient intake, host genetics and ethnicity, lifestyle, and host defence mechanisms (Marsh and Devine, 2011).

Before metagenomics techniques were introduced in the 2000s, the oral bacteria were only characterised by culturing them in the laboratory. As is estimated that only $\sim 50\%$ of bacteria in the mouth can be cultured (Wade, 2013), many bacteria could previously not be studied. Gibbons and Houte (1975) summarised 9 genera characterised at the time, which covered five phyla: Firmicutes, Bacteroidetes, Proteobacteria, Spirochaetes, and Fusobacterium. In more recent years, around 600-700 (mostly commensal, i.e. non-pathogenic) bacterial taxa covering 13 phyla have been characterised (Chen *et al.*, 2010; Dewhirst *et al.*, 2010; Jenkinson, 2011), with some studies suggesting that the overall number of species could increase to approximately 1200 due to continuing improvements in technology (Jenkinson, 2011). This high level of species diversity is likely to be a result of many different microenvironments within the mouth that are subject to different conditions, e.g. the tongue, palate, cheeks, mucosal glands, gums and teeth. Metagenomic analyses have even shown that biofilms on different teeth of the same person can differ in their microbial composition (Shestakov, 2011).

The bacteria in the human mouth can obtain nutritional substrates from the host, the diet, and from other bacteria. Nutrients derived from the host include constituents of saliva, desquamated (i.e. scraped off) epithelial cells, and serum-like

crevicular fluid, which is secreted from the parotid, submaxillary, and minor glands, all found in different areas of mouth (Gibbons and Houte, 1975). It is thought that each gland can exert local selective pressures on bacterial biofilm development.

Ingested food is only available intermittently to bacteria in the mouth, however insoluble fibres and adhesive foods may be retained in the mouth if they are trapped. While little is known about the effects of dietary proteins and lipids on oral bacteria, carbohydrates have been studied due to the link of their breakdown, by Streptococcus mutans, with tooth decay. Both saliva and crevicular fluid contain low levels of glucose and free amino acids, as well as glycoproteins. Bacterial communities have been shown to work together to degrade complex glycoproteins as single bacterium members do not have the capability (Wickström *et al.*, 2009).

Within the oral microbiome there are a number of bacterial species shown to play a key role in oral disease, mainly in dental caries and periodontal diseases (Marsh and Devine, 2011; Wade, 2013). Dental caries, or tooth decay, results from a high-carbohydrate diet. Acid-producing bacteria, particularly *Streptococcus mutans* and Lactobacilli, produce acid more frequently under such conditions, damaging dental tissue and producing a more acidic environment which promotes the adhesion of similar tooth decay-related bacteria (Wade, 2013). Periodontal diseases include gingivitis and periodontis. Both of these diseases are typically caused by anaerobic Gram-negative bacteria found in subgingival plaque, which is located below the gum line (Belda-Ferre *et al.*, 2011).

4.1.10 Aims and objectives

The work reported in this chapter uses computational methods to classify bacterial organisms and provide functional annotations for their proteins, within three human oral metagenome data sets. These samples have been sourced from the tongue, dental plaque and mucosal glands in the oral cavity.

The main aims are to: 1) build a protocol that can be used in the analysis of future metagenomic data sets, 2) annotate the metagenomes with functional infor-

mation, and 3) compare the functional profiles between different oral environments, and between the oral and gut microbiomes.

4.2 Methods

4.2.1 Data sets

Human tongue data set Bacterial DNA sampling, extraction and preparation for the human tongue metagenome data set was undertaken as described in Easton (2009). To summarise, 9 volunteers from the Ward Group in University College London brushed their tongue with a toothbrush for one minute, twice a week for 3 or 4 weeks. The samples were immediately frozen, and the DNA later extracted and pooled ready for sequencing.

Human dental plaque and mucosal glands data sets Clinical samples were collected from 27 patients at the Kings College London Dental Institute that presented various stages of periodontal disease. In each patient, both supra- and subgingival plaque was collected from all teeth in a randomly selected area (Hunter *et al.*, 2011). Samples were also taken from all oral mucosal glands. The metagenomic DNA was extracted from the 27 individual dental plaque and mucosal glands samples and then pooled in preparation for sequencing.

Human gut data set The healthy gut data set contains faecal samples taken from 110 individuals across the world (Yatsunenko *et al.*, 2012) including children and adults of the Amazonas of Venezuela, rural Malawi, and US metropolitan areas. The whole genome shotgun faecal metagenome data was sequenced using 454 pyrosequencing and is publicly available from the MG-RAST metagenomics analysis server (`http://metagenomics.anl.gov/linkin.cgi?project=98`) (Yatsunenko *et al.*, 2012).

A representative bacterial genomes data set A total of 1438 representative bacterial genomes were downloaded from the Representative Proteomes project (Chen *et al.*, 2011). This set of representatives have been selected by Chen *et al.* (2011) from the Protein Information Resource to reduce the bacterial genome sequence

space as much as possible, whilst maintaining the functional annotation and sequence diversity. This data set was used for comparing the functional repertoires of the metagenomes studied in this chapter, with a background bacterial set of genomes.

4.2.2 DNA sequencing

The DNA samples for each oral environment were combined to give a pooled sample. Each of the three pooled samples was then sequenced using 454 pyrosequencing. The human tongue DNA was sequenced by Tony Brooks at the UCL Eastman Dental Institute (London, UK). The human dental plaque DNA and the mucosal gland DNA was sequenced by Alan Walker at the Sanger Institute (Cambridge, UK). All reads were sequenced using the GS FLX System for DNA sequencing (454 Life Sciences Corporation).

4.2.3 Generation and processing of metagenome sequence data

4.2.3.1 Generating FASTA sequence data

Following sequencing, the raw image data was converted into FASTA format through image processing and signal processing. Image processing generated the Composite Well Files (CWFs), which contain the normalised signal value for each well over each flow. Signal processing was then performed to correct for known errors; this involved data filtering based on signal quality, trimming read ends for low quality and primer sequences, and the generation of Standard Flowgram Format (SFF) files, which are used as standard to encode pyrosequencing results from 454 sequencing. The FASTA files were extracted from these SFF files. These steps were performed by Tony Brooks for the tongue data set, and by Alan Walker for the dental plaque and mucosal glands data sets.

4.2.3.2 Quality assessment of sequence reads

To assess the quality of the sequence reads in each data set, the sequence (FASTA) and the quality (QUAL) files were first converted into FASTQ format. Sequence data statistics were generated for sequence length, GC content, quality scores, n-plicates, complexity, tag sequences, poly-A/T tails and odds ratios. This was done using the PReprocessing and INformation of SEQuences (PRINSEQ) tool (online version 0.16.1 beta, `http://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi`). This tool is explicitly designed for 454/Roche data. There is a report on the data quality, which is then used to filter, reformat and trim the sequence data (Schmieder and Edwards, 2011b).

The data were then stringently filtered using parameters suggested by PRINSEQ (see Table 4.1). The quality scores for 454 sequences are Phred-based and range from 0 to 40 (Schmieder and Edwards, 2011b). Bases with low Phred quality scores on the end of sequence reads were trimmed off, and sequence reads with a mean Phred quality score less than 20 were removed. Reads that are too short or too long were removed. Very short sequence reads can cause problems. For example when using them to search for homologous proteins in a database they are more likely to match a sequence by chance. On the other hand, very long sequence reads are likely to contain long stretches of homopolymer runs, which is a reported issue of pyrosequencing technologies (Huse *et al.*, 2007). The presence of ambiguous bases, 'N's, in a sequence read can indicate a low quality sequence, therefore any sequences with more than 1% N content were removed. Low complexity sequence reads are those containing stretches of bases with little information content, such as di- and trinucleotide repeats (Schmieder and Edwards, 2011b). An entropy approach was used to evaluate the entropy of trinucleotides in a sequence; entropy values range from 0 to 100, where 0 represent a sequence containing a single amino acid and 100 represents a sequence containing no repeats. Finally, the filtering of sequence read duplicates was performed to remove any PCR amplification bias introduced before sequencing (Schmieder and Edwards, 2011b).

	Parameters applied
Trim low quality ends	Mean $\geq$ 15, W:2, S:1
Filter too short and/or too long reads	Mode $\pm$ 2 SD
Filter low quality reads	Mean $\geq$ 20
Filter reads with ambiguous base N	$\leq$ 1% (1 out of 100)
Filter low complexity reads	Entropy $\geq$ 50
Filter read duplicates	5' and reverse complement

Table 4.1: A description of the parameters defined to filter the human oral metagenome sequence reads using the PRINSEQ web tool. W = window size, S = step size.

4.2.3.3 Detection and removal of human contamination

Good-quality sequences were compared against the human reference genome build 37 to detect human contamination. The human reference genome was first downloaded from the NCBI FTP server, extracted and filtered using PRINSEQ to remove ambiguous bases, short sequences and sequence duplicates. The data was then converted into database format. This process was carried out as described in the DeconSeq online manual `http://deconseq.sourceforge.net/manual.html`.

Sequences were scanned against the human reference genome database with DeconSeq (Decontamination of sequence data) standalone version 0.4.1 (Schmieder and Edwards, 2011a). The BWA-SW algorithm which is used by DeconSeq to identify human contaminant sequences has been reported to be ten times faster than BLAST (Li and Durbin, 2010). Sequences were identified as contaminants if they matched a human sequence with $\geqslant$ 90% sequence identity and an alignment coverage of $\geqslant$ 90%. The data was then converted back into FASTA format for analysis.

4.2.4 Characterising bacterial species in the human oral microbiome

To classify the bacterial species in each data set, each set of filtered sequence reads were scanned against the 16S rRNA RDP database via the MG-RAST server. Annotations were only considered with E-values $\leqslant$ 1x10^{-5}, sequence identities $\geqslant$ 99%,

and a minimum alignment length of 50bp.

Bacterial phyla were classified in each data set by scanning the filtered sequence reads against the 16S rRNA RDP database using the RDP standalone version 2.9. A confidence threshold of 0.8 was used to accept a classification.

4.2.4.1 Analysing GC content

Metagenome analyses frequently combine the results of multiple methods to examine the species in the microbiome. Another method of analysing the bacterial species is through the examination of sequence read GC content. The fractional GC content of each sequence in the data set was calculated using the program geecee, by EMBOSS version 6.3.1 (Rice *et al.*, 2000). The normalmixEM function from the R package, mixtools (Benaglia *et al.*, 2009) was used to estimate whether there was a mixture of normal distributions present in each data set as it is common for metagenome data to contain a bimodal distribution (Schmieder and Edwards, 2011b).

4.2.5 Assembling metagenome sequence data

4.2.5.1 Assembly of sequence reads

To generate contiguous sequences, or contigs, the data was assembled *de novo* using Mimicking Intelligent Read Assembly (MIRA) version 3.2.1, a whole genome shotgun and EST sequence assembler for Sanger, 454, Solexa (Illumina), IonTorrent and PacBio data (Chevreux *et al.*, 2004). There are less *de novo* assembly tools available for 454 pyrosequencing data in comparison to Illumina data, and MIRA has been recommended in the literature (Hooper *et al.*, 2010; Dutilh *et al.*, 2012; Lysholm *et al.*, 2012; Tse *et al.*, 2012). Muhammadzadeh *et al.* (2013) found MIRA to perform slightly better than other comparable *de novo* assemblers.

Only sequence reads with more than 50 bases were considered for assembly. Spoiler detection was switched on, which typically prevents chimeric contigs from joining.

4.2.5.2 Gene prediction

Genes were predicted within the contigs using the MetaGeneMark standalone program (Zhu *et al.*, 2010). The program is based upon GeneMark, which identifies genes in complete genomes, and has been adapted to predict genes in short fragments. Codon and oligomer frequencies are used to build heuristic models, which are then used to find genes in the short reads. Trimble *et al.* (2012) evaluated five popular *ab initio* gene prediction algorithms, FragGeneScan, MetaGeneAnnotator, MetaGeneMark, Orphelia, and Prodigal. They compared the performance of these methods over different rates of simulated sequencing error and also on metagenomic datasets. They concluded that MGM offers accurate gene predictions, as long as the sequence data is error-free, and that it is best-suited for gene prediction in higher-quality sequences such as assembled contigs. To increase the chances of finding full-length genes and to provide the most accurate prediction possible, the prediction tool was run using the contig data sets.

4.2.6 Characterising bacterial protein function in the human microbiomes

Predicting protein function is a difficult task. Multiple methods have therefore been used in this stage of the analysis to obtain more reliable answers. The metagenome sequence fragments were first searched against entire protein sequences using the MG-RAST server. Subsequently, since the fragments are generally shorter than a full-length protein, in house CATH, domain-based methods, DomainFinder 3 and FunFHMMER were used as they were more likely to identify domains in the shorter fragments.

4.2.6.1 Whole protein-based prediction

Sequence reads from the three oral data sets were searched against the SEED database and categorised into Subsystems functional groups through the MG-RAST

server. Annotations were only considered if they had E-values $\leqslant 1\text{x}10^{-5}$, sequence identities $\geqslant 60\%$, and a minimum alignment length of 50bp.

4.2.6.2 Domain-based prediction

DomainFinder3 is an algorithm that uses a hidden Markov model (HMM) approach to detect CATH protein domain families in a query sequence (Yeats *et al.*, 2010). The filtered 454-sequenced tongue metagenome DNA sequence reads were first converted into their respective 6-frame translated protein sequence using the tool 'transeq' from EMBOSS-6.3.1 (Rice *et al.*, 2000). These protein sequence fragments were submitted to DomainFinder3, which scanned these sequences against CATH and Pfam-A HMMs using HMMER3 software (Mistry *et al.*, 2013) to identify CATH structural domains and Pfam domains. The CATH HMMs were constructed from the CATH v4.0 library. Pfam-A HMMs were used from version 27.0 of the Pfam database.

Once a domain had been assigned to a superfamily it was rescanned against all the functional family HMMs for that superfamily. This retrieved structural and functional information about each domain. The Pfam functional families were used to provide annotations for domains which could not be mapped to CATH.

The FunFHMMER method in this chapter is an update to the Funfamer and DFX methods used in the previous work chapters. All three methods use the GeMMA method to perform an initial profile-based clustering of protein domain sequences. FunFHMMER decides whether two clusters should be merged or kept separate depending on whether there is sufficient similarity in specificity determining residues (SDPs) between the clusters. This method uses the GroupSim approach of Capra and Singh (2008) to detect specificity determining residues. FunFHMMER is the most recent version of the function prediction method based on CATH functional families (Figure A.1) and has been shown to outperform FunFamer and DFX (Sayoni Das, personal communication). It was ranked second for the prediction of GO terms in the Biological Process ontology and fourth for the prediction

of GO terms in the Molecular Function ontology (out of 110 methods) in the recent CAFA-2 function prediction method.

4.2.7 Comparing the functional profiles of oral and gut microbiomes

To explore whether the bacterial communities in the healthy tongue and gut metagenome data were significantly abundant in certain types of protein function, they were first each compared against a set of representative bacterial genomes. This background was used as, in some cases, two metagenomes may have a number of enriched genes in common, which would be missed if they were simply compared against each other. Through the comparison of each metagenome with an average bacterial background, such enrichments will not be missed.

The tongue and gut metagenome data were also compared against each other to explore the differences in functional repertoires. Functional family profiles and KO term profiles have been compared.

4.2.7.1 Protocol to identify significant changes in FunFam abundance between data

To calculate FunFam enrichment between the tongue and gut metagenomes, and between each of these two metagenomes and the bacterial background, the number of sequences assigned to each FunFam was counted together with the number of sequences not assigned to all other FunFams. The Fisher exact test used these counts to determine whether there was functional enrichment of a particular FunFam with a confidence value threshold of 0.95.

4.2.7.2 Identifying enriched metabolic genes and pathways in the tongue and gut data

To investigate whether there was significant enrichment of metabolic genes in the tongue and gut metagenomes, sequence reads in both metagenomes were functionally

annotated with KEGG Orthology (KO) terms through the MG-RAST server (Meyer *et al.*, 2008).

All KO annotations were downloaded through the MG-RAST API and each data set was filtered to remove poor quality functional assignments, e.g. possible misannotations resulting from assignments made to short read lengths. Annotations were only considered with E-values $\leqslant 1\text{x}10^{-5}$, sequence identities $\geqslant 75\%$, and a minimum alignment length of 50bp.

For each KO term in the tongue and gut filtered data sets, the number of sequence reads assigned and not assigned was counted. The Fisher exact test used these counts to calculate functional enrichment of each KO term with a confidence value threshold of 0.95.

The pathway enrichment score (developed by Enav *et al.* (2014)) was used to calculate enrichment of the pathways containing enriched KO terms. The enrichment score is defined in Equation 4.2:

$$PES = \frac{N * \sum_{N}^{1}(-log_{10}(Pn))}{T} \tag{4.2}$$

PES represents the pathway enrichment score, N is the number of enriched KO terms in a particular pathway for a given data set, Pn is the p-value calculated by the Fisher exact test for each enriched KO term, and T is the total number of KO terms in a particular KEGG pathway.

4.3 Results

Whole genome shotgun metagenome data from three human oral data sets (dental plaque, mucosal glands, and tongue) have been processed and characterised using multiple computational methods to classify the bacterial species and to predict protein function annotations and enrichments.

Publicly-available metagenome data for a set of 110 healthy children and adults have been compared with the healthy tongue metagenome data to identify sets of metabolic pathways differentially enriched in the tongue and the gut microbiomes.

4.3.1 Processing of sequence data

Table 4.2 reports the number of raw 454 metagenome DNA sequence reads produced from the pooled human oral samples. The three data sets contained a similar number of sequence reads.

Data set	**Sequence reads**	**Base pairs**	**Pooled samples**
Dental plaque	1,364,333	492,833,154	27
Mucosal glands	1,210,767	384,156,658	27
Tongue	1,182,079	455,966,019	9

Table 4.2: DNA sequence reads sequenced from pooled human oral metagenome samples using 454 pyrosequencing methods.

4.3.1.1 Quality assessment of sequence reads

Following the filtering and trimming of sequence reads with the PRINSEQ tool, $\sim$ 10-15% of sequence reads were removed from each data set due to their poor quality (see Table 4.3). Low quality refers to: sequences with a low average Phred quality scores, bases with low Phred quality scores, reads that were too short or too long, reads containing ambiguous bases, reads with long low complexity regions, and exact duplicates of reads.

Data set	**# from original sample (%)**
Dental plaque	1,201,749 (88.08 %)
Mucosal glands	1,046,288 (86.42 %)
Tongue	1,061,991 (89.84 %)

Table 4.3: Good quality DNA sequence reads remaining after filtering.

4.3.1.2 Detection and removal of human contamination

Sequence reads were scanned against the human reference genome build 37 using BLAST to detect human contamination. Table 4.4 lists the number of sequence reads remaining in each data set following these quality checks. A high percentage of the mucosal sequence reads were removed with this step.

Data set	**# from original sample (%)**
Dental plaque	946,343 (78.75 %)
Mucosal glands	246,758 (23.58 %)
Tongue	985,846 (92.83 %)

Table 4.4: DNA sequence reads remaining after the removal of sequence reads identified as human contamination.

Figure 4.1 summarises the number of reads filtered out throughout the different stages of quality processing.

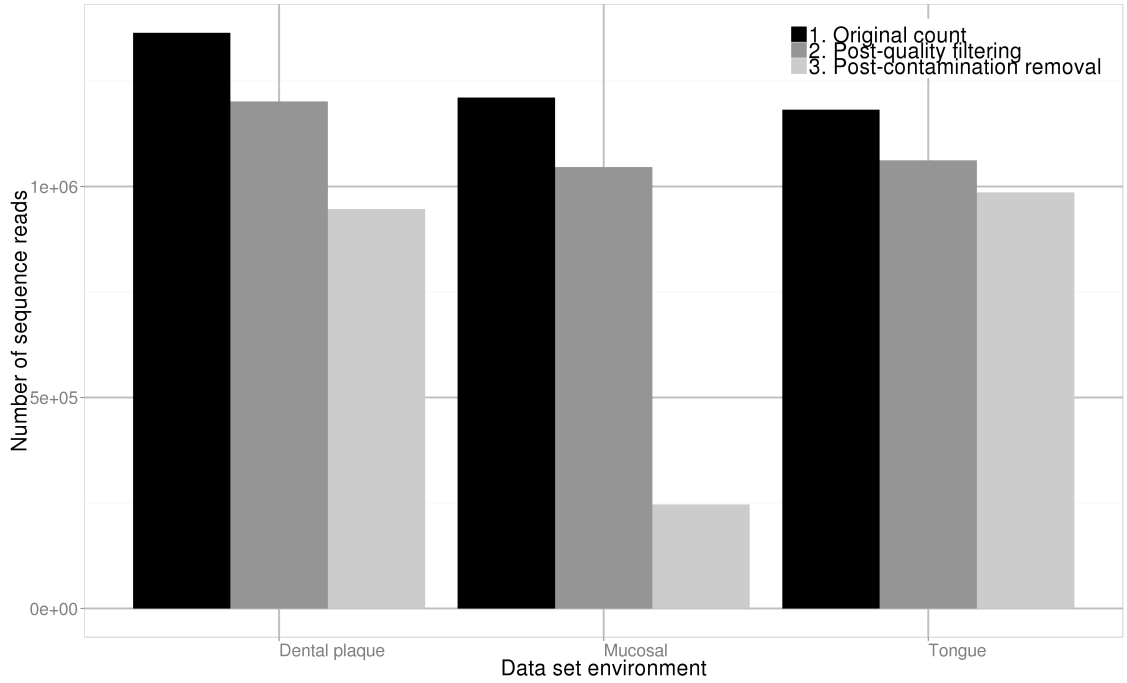


Figure 4.1: The number of sequence reads remaining after two stages of filtering. Step 2 represents the filtering performed with the PRINSEQ tool. Step 3 represents the filtering performed with the DeconSeq tool.

4.3.2 Characterising bacterial species in the human oral microbiome

Table 4.5 shows the bacterial phyla characterised within each data set. This is the result of scanning the sequence reads from each data set against the 16S ribosomal RNA (rRNA) gene sequences in the Ribosomal Data Project (RDP) database. A total of 1.8% (16,984 out of 946,343), 2.0% (4,903 out of 246,758), and 0.7% (7,006 out of 985,846) of the filtered sequence reads mapped to 16S rRNA genes in the RDP database from the dental plaque, mucosal glands, and tongue data sets, respectively.

A total of 284, 137, and 195 bacterial strains were identified in the dental plaque, mucosal glands, and tongue metagenome data, respectively. In comparison with previous studies that have typically found around 500 different bacterial species in the mouth, these numbers appear rather low. This may be due to issues with DNA extraction from some bacterial species or because the 454 sequencing technique provided low coverage for certain bacterial species. The sequence reads were

also scanned against the Greengenes rRNA database as well as the large subunit (LSU) and small subunit (SSU) rRNA databases from SILVA via MG-RAST to determine whether the number of bacterial strains was low because of biases in the RDP database searched. These additional three databases however provided similar results. The RDP scan results publicly available through MG-RAST report that the gut metagenome on the other hand has a total of 1043 different bacterial strains, which is around the number suggested by the literature.

Table 4.5 reports that a total of 10, 9, and 10 bacterial phyla were classified in the dental plaque, mucosal glands, and tongue data sets, respectively. The Firmicutes, and Proteobacteria are the two most abundant phyla in all three metagenomes (highlighted in bold italics), with Firmicutes being most abundant in the mucosal glands data and Proteobacteria in the dental plaque and tongue data. The Actinobacteria, Bacteroidetes, and Fusobacteria phyla are less abundant, however still make up approximately % of the sequences classified. The Aquificae and Verrucomicrobia, the SR1, and the Spirochaetes and SR1 phyla re the least abundant in the dental plaque, mucosal glands, and tongue data sets, respectively.

Phylum	**Dental plaque**	**Mucosal glands**	**Tongue**
Actinobacteria	189 (1.11%)	114 (2.33%)	566 (8.08%)
Aquificae	1 (0.01%)	0	0
Bacteroidetes	433 (2.55%)	88 (1.79%)	443 (6.32%)
Candidatus Saccharibacteria	114 (0.67%)	4 (0.08%)	15 (0.21%)
Chloroflexi	0	0	5 (0.07%)
Cyanobacteria/ Chloroplast	2 (0.01%)	2 (0.04%)	2 (0.03%)
Firmicutes	***1788 (10.53%)***	***3743 (76.34%)***	***1325 (18.91%)***
Fusobacteria	456 (2.68%)	14 (0.29%)	72 (1.03%)
Proteobacteria	***13925 (81.99%)***	***937 (19.11%)***	***4576 (65.32%)***
Spirochaetes	75 (0.44%)	0	1 (0.01%)
SR1	0	1 (0.02%)	1 (0.01%)
Verrucomicrobia	1 (0.01%)	0	0
Total 16S rRNA sequences	16984	4903	7006

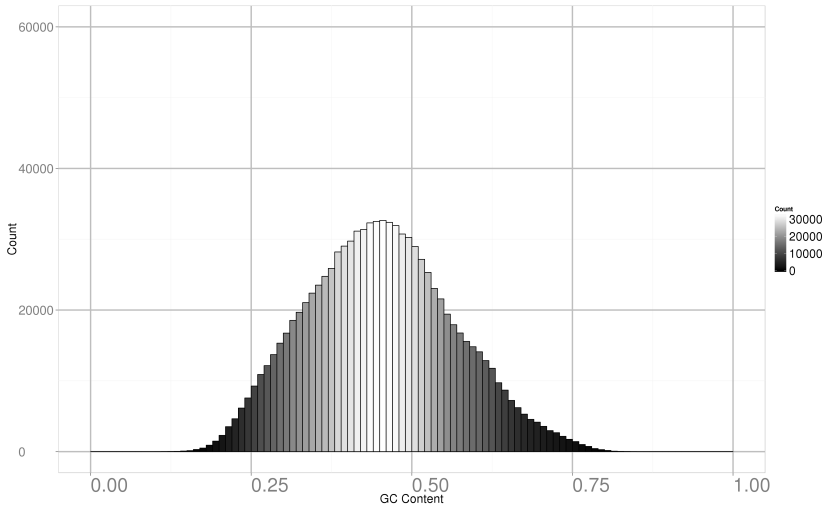
Table 4.5: Number of sequence reads classified to bacterial phyla identified in the three human oral metagenome data sets. Percentages represent the proportion of metagenome sequences classified to a phyla out of the proportion of all 16S rRNA metagenome sequences identified (see final row). The most dominant phyla, Firmicutes and Proteobacteria, are highlighted in bold italics.

These results can be compared with species analysis of the human gut previously undertaken by Kaoutari *et al.* (2013) who constructed a 'mini-microbiome' to represent the bacterial phyla found in the typical healthy adult gut microbiome. Their data came from five different metagenomic studies and consisted of 177 genomes across 12 phyla. As in the oral microbiome data presented in Table 4.5, the Firmicutes phyla was found to be among the most abundant with 104 (out of 177, $\sim 59\%$) genomes reported. While Table 4.5 reports that bacteria from the Proteobacteria are the most abundant in the Dental plaque and Tongue environments,Kaoutari *et al.* (2013) only found 22 members (out of 177, $\sim$ 12%) from this phyla in their gut metagenome data. Table 4.5 shows that bacteria from the Actinobacteria and Bacteroidetes phyla are less abundant but still contribute between 1-8% of sequence reads, depending on the oral environment studied. Kaoutari *et al.* (2013) also found the Actinobacteria and Bacteroidetes phyla to be fairly abundant in the gut, with 12 and 29 genomes reported, respectively.

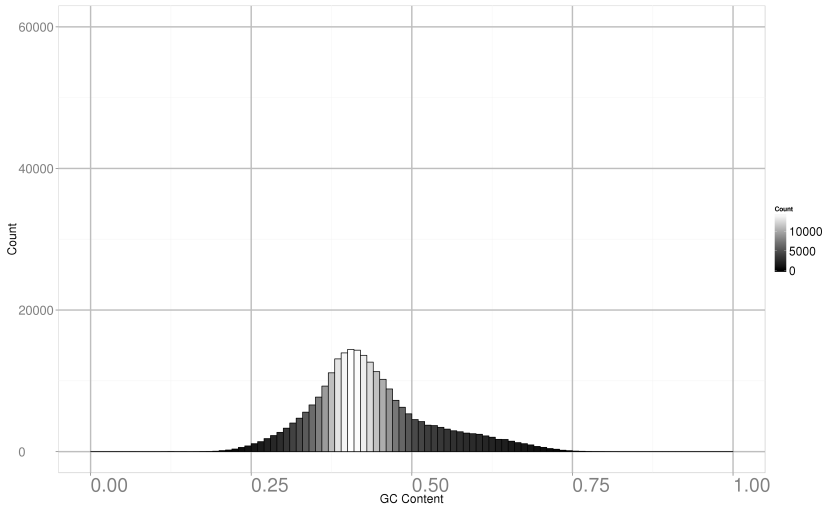
4.3.2.1 Analysis of sequence GC content

Figure 4.2 describes the GC content distribution of the three oral data sets. All three data sets have a clear, major peak with mode values of approximately 0.4 (i.e. a 40% GC content), which is probably due to an abundance of Firmicutes and Bacteroidetes as these two phyla typically have a low GC content.

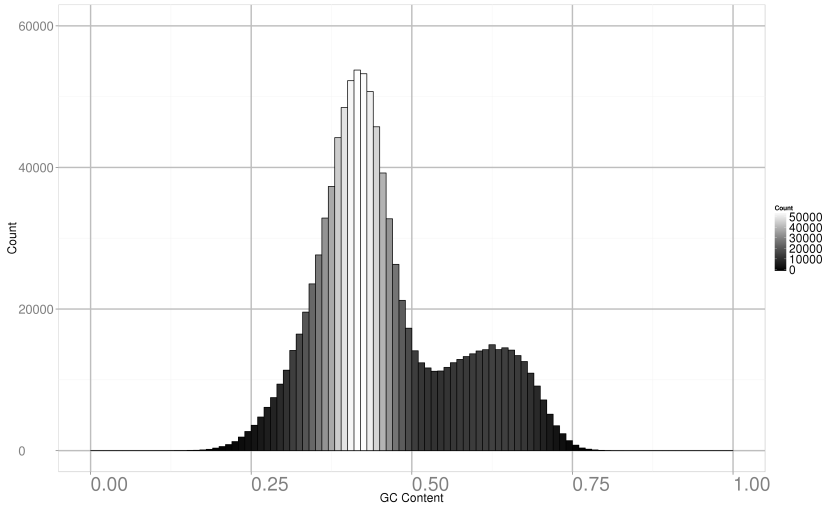
As the distributions in Figure 4.2 appear to be bimodal, the normalmixEM function in R was used to examine this. Figure 4.3 shows two distributions that have been fit separately to the mucosal glands and the tongue data. This suggests a second, minor peak with mode values at 0.594 and 0.625 in the mucosal glands and the tongue data, respectively. This corresponds to an abundance of Proteobacteria, which have genomic GC contents ranging from 30-60% (Hildebrand *et al.*, 2010).



(a) Dental plaque GC content distribution

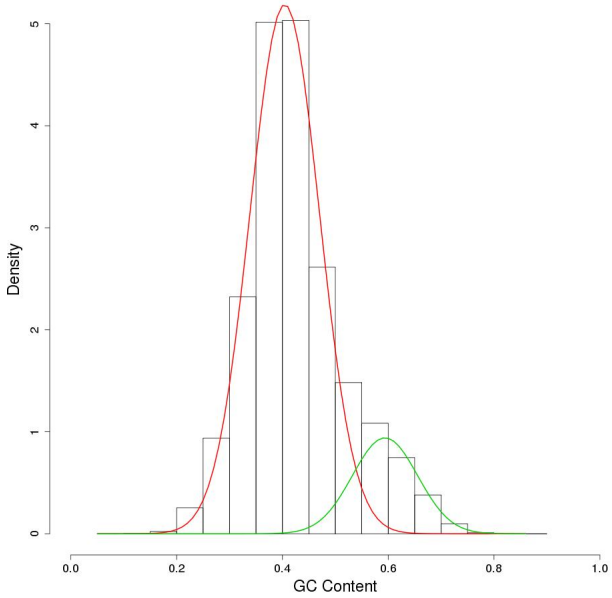


(b) Mucosal glands GC content distribution

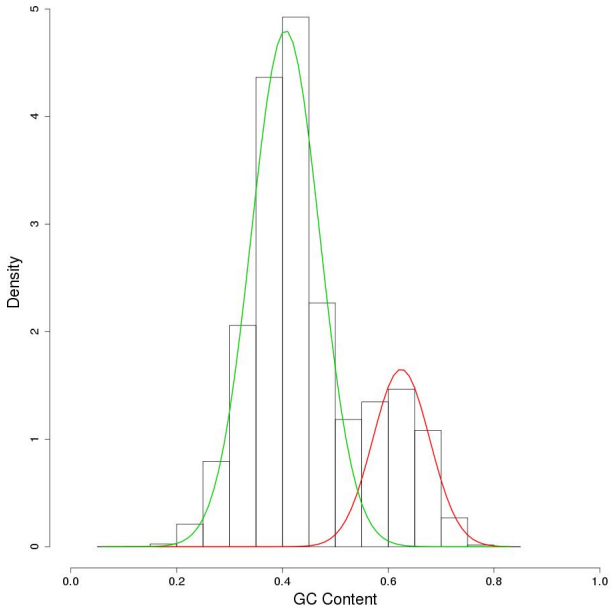


(c) Tongue GC content distribution

Figure 4.2: Fractional GC content distribution of the human oral metagenomes.



(a) Mucosal GC content bimodal distribution



(b) Tongue GC content bimodal distribution

Figure 4.3: The distribution of fractional GC content across tongue and mucosal sequence reads. The normalmixEM function from the R package, mixtools identified a bimodal distribution in the two data sets. The modes in the tongue data were calculated as 0.406 and 0.625. The modes in mucosal data were calculated as 0.404 and 0.594.

4.3.3 Assembling the metagenome sequence reads

4.3.3.1 Assembling the human oral metagenome sequence reads

All three oral data sets were assembled using MIRA software. The general statistics of the resulting contiguous reads, or contigs, are summarised in Table 4.6.

The mucosal glands data set has had fewer reads assembled compared to the dental plaque and tongue data sets due to the smaller number of reads remaining after filtering. The N50 contig size (see definition in Section 4.1.7, page 172) is largest for the tongue data, which also has the longest contig length.

Data set	**Sequence reads assembled**	**Contigs**	**N50 contig size (bp)**	**Longest contig length (bp)**
Dental plaque	421,565	34,842	1,423	31,355
Mucosal	96,067	8,312	1,346	11,691
Tongue	657,898	39,971	2,396	41,967

Table 4.6: General statistics from the contiguous sequences assembled.

4.3.3.2 Gene prediction

Full-length genes have been predicted within the contig data using MetaGeneMark. Table 4.7 lists the number of predicted genes in each of the data sets.

Data set	**# Contigs**	**# Genes**
Dental plaque	34,842	79,135
Mucosal	8,312	18,227
Tongue	39,971	125,211

Table 4.7: The number of genes predicted from the assembled contigs.

4.3.4 Functional annotation using the MG-RAST web server

The three oral data sets were scanned against the Subsystems database to assign functional annotations. Subsystems represents four hierarchy levels of annotation,

Figure 4.4 shows the percentage of sequence read assignments to the top hierarchical level. Subsystems functional annotations were assigned to 463,941 plaque sequence reads (49.02%), 145,187 mucosal sequence reads (58.84%), and 594,344 tongue sequence reads (60.29%). The most populated category is clustering-based subsystems, which represents groups of protein sequences with unknown function. Carbohydrate metabolism and protein metabolism are the next most populated categories, which is likely to be a result of bacteria breaking down carbohydrates and glycoproteins for nutrients. The least populated category is photosynthesis, which is likely to be a result of a small number of photosynthetic bacteria present that have some small amount of exposure to sunlight. All three data sets have similar percentage assignments to each category.

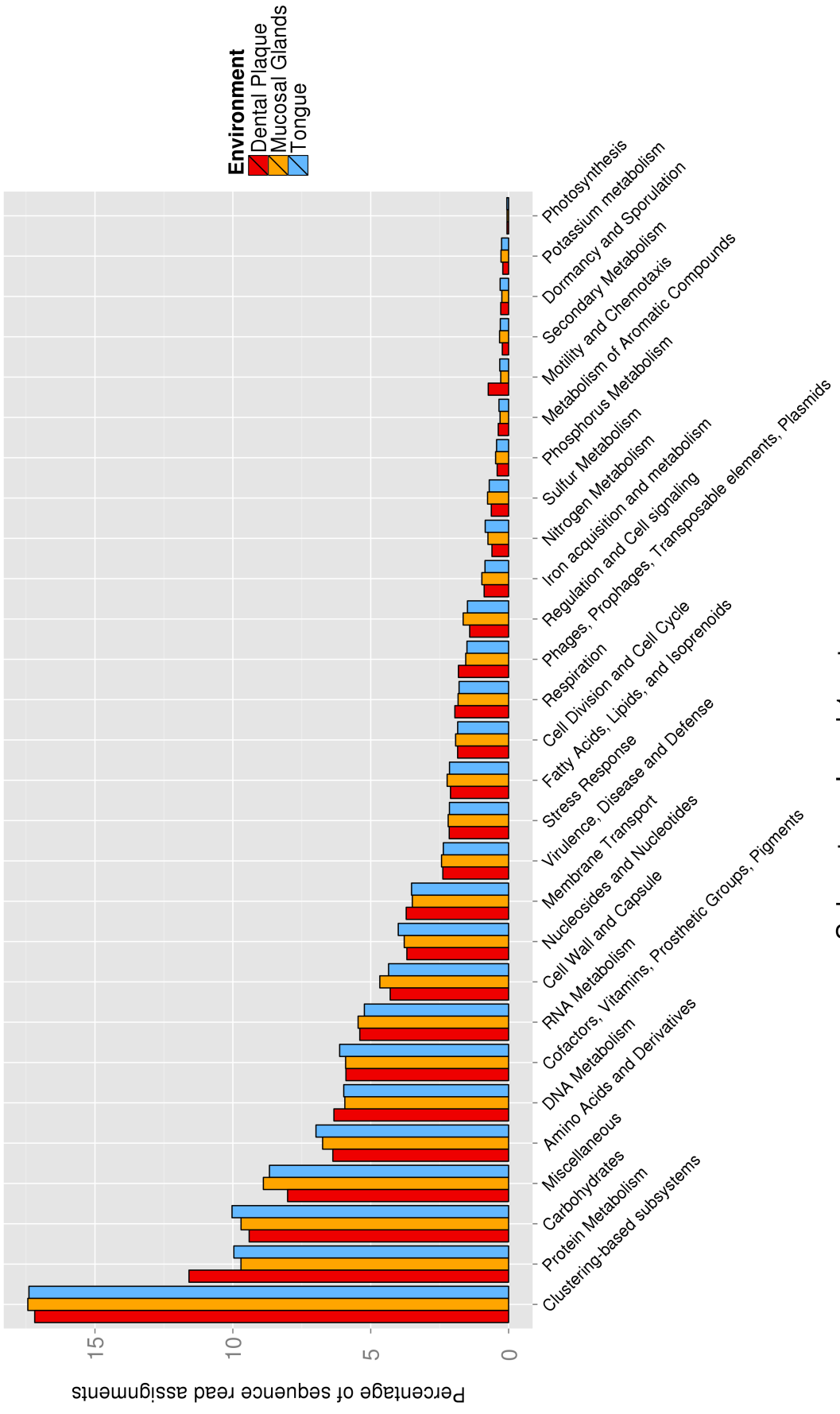


Figure 4.4: The percentage of functionally-annotated oral metagenome sequence reads associated with the 28 top-level Subsystems functional categories.

4.3.5 Functional annotation using CATH and Pfam functional families

4.3.5.1 Coverage statistics

Metagenome sequences reads, contigs, and genes were assigned to CATH and Pfam protein domain homologous superfamilies in CATH version 4.0 and Pfam version 27. CATH version 4.0 comprises 2,734 superfamilies and 110,439 functional families. Pfam version 27 comprises 14,831 superfamilies and 95,027 functional families.

To calculate the coverage attained through mapping metagenome data to CATH and Pfam superfamilies and functional families, two general statistics were used: 1) the number of different superfamilies and different functional families assigned to the data, and 2) the number of different sequences (either read, contig, or gene sequences) that were assigned to superfamilies and functional families (Tables 4.8 and 4.9).

	# CATH superfamilies assigned	# sequences assigned to superfamilies	# CATH FunFams assigned	# sequences assigned to FunFams
Dental plaque reads	1625 (/2734 59.44%)	478485 (/5678058, 8.43%)	3098 (/110439, 2.81%)	43062 (/ 5678058, 0.76%)
Mucosal reads	1438 (/2734 52.60%)	116368 (/1480548, 7.86%)	2279 (/110439, 2.06%)	12897 (/ 1480548, 0.87%)
Tongue reads	1526 (/2734 59.44%)	540250 (/5915076, 9.13%)	3142 (/110439, 2.85%)	57511 (/5915076, 0.97%)
Dental plaque contigs	1291 (/2734 47.22%)	35194 (/209052, 16.84%)	2339 (/110439, 2.12%)	8226 (/ 209052, 3.93%)
Mucosal contigs	1089 (/2734 39.83%)	8232 (/49872, 16.51%)	1614 (/110439, 1.46%)	2364 (/ 49872, 4.74%)
Tongue contigs	1336 (/2734 48.87%)	54959 (/239826, 22.9%)	3214 (/110439, 2.91%)	16015 (/239826, 6.68%)
Dental plaque genes	1282 (/2734 46.89%)	37281 (/79135, 47.11%)	2194 (/110439, 1.99%)	8029 (/ 79135, 10.15%)
Mucosal genes	1076 (/2734 39.36%)	8687 (/18227, 47.66%)	1477 (/110439, 1.34%)	2280 (/ 18227, 12.51%)
Tongue genes	1335 (/2734 48.83%)	65557 (/125211, 52.36%)	3047 (/110439, 2.76%)	16825 (/125211, 13.44%)

Table 4.8: Coverage statistics on sequence assignments to CATH superfamily and functional family groups.

	# Pfam superfamilies assigned	# sequences assigned to superfamilies	# Pfam FunFams assigned	# sequences assigned to FunFams
Dental plaque reads	5303 (/14831 35.76%)	504756 (/5678058, 8.89%)	5113 (/95027, 5.38%)	59989 (/5678058, 1.06%)
Mucosal reads	4254 (/14831 28.68%)	126246 (/1480548, 8.53%)	3229 (/95027, 3.40%)	17475 (/1480548, 1.18%)
Tongue reads	4827 (/14831 32.55%)	570044 (/5915076, 9.64%)	4740 (/95027, 4.99%)	77727 (/5915076, 1.31%)
Dental plaque contigs	3634 (/14831 24.50%)	39577 (/209052, 18.93%)	3417 (/95027, 3.60%)	11213 (/209052, 5.36%)
Mucosal contigs	2574 (/14831 17.36%)	9380 (/49872, 18.81%)	2012 (/95027, 2.12%)	3097 (/49872, 6.21%)
Tongue contigs	3768 (/14831 25.41%)	61230 (/239826, 25.53%)	4410 (/95027, 4.64%)	21060 (/239826, 8.78%)
Dental plaque genes	3600 (/14831 24.27%)	42030 (/79135, 53.11%)	3209 (/95027, 3.38%)	10348 (/79135, 13.08%)
Mucosal genes	2520 (/14831 16.99%)	9767 (/18227, 53.59%)	1816 (/95027, 1.91%)	2866 (/18227, 15.72%)
Tongue genes	3785 (/14831 25.52%)	74356 (/125211, 59.38%)	4197 (/95027, 4.42%)	21595 (/125211, 17.25%)

Table 4.9: Coverage statistics on sequence assignments to Pfam superfamily and functional family groups.

Tables 4.8 and 4.9 report the number of metagenome sequences assigned to CATH and Pfam superfamilies and functional families. Almost twice as many CATH superfamilies have been assigned, compared to Pfam superfamily assignments. However even so, less than 10% of sequence reads are assigned to a superfamily. This is likely to be due to the reads being very small ($\sim$ 450bp) and this may mean that some matches do not meet the threshold for superfamily recognition.

Unsurprisingly, when moving from the short sequence reads, to contigs and genes (Tables 4.8 and 4.9), a larger amount of the data is assigned to both CATH and Pfam FunFams.

Pfam FunFams were used to provide additional functional annotations where CATH could not.

Due to the short nature of the sequence reads, we wanted to investigate the number of domains mapped to a read and how this number increased when the reads were assembled into contigs. Figure 4.5 shows that the sequence reads typically contain matches to one or two CATH or Pfam domains, whereas the contigs are shown to contain more domains, as expected.

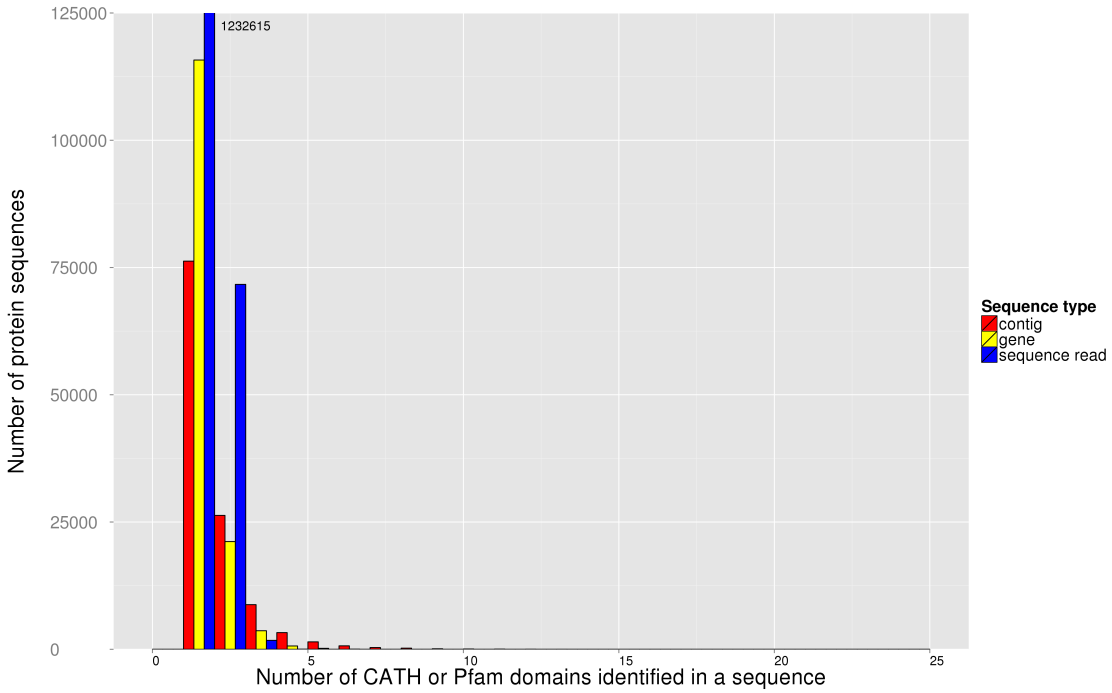


Figure 4.5: The number of domains identified in the tongue metagenome per read, per contig, and per predicted gene. Each sequence can have one or more CATH or Pfam domain identified.

To determine whether metagenome sequences are matching complete domains or partial domains, the percentage coverage of each domain HMM match to its query was calculated. Figure 4.6 shows that the percentage coverage of an identified domain increased with length of the sequence type, i.e. a sequence read had a median coverage of 42.3%, a gene had a median coverage of 68.9%, and a contig had a median coverage of 70.8%.

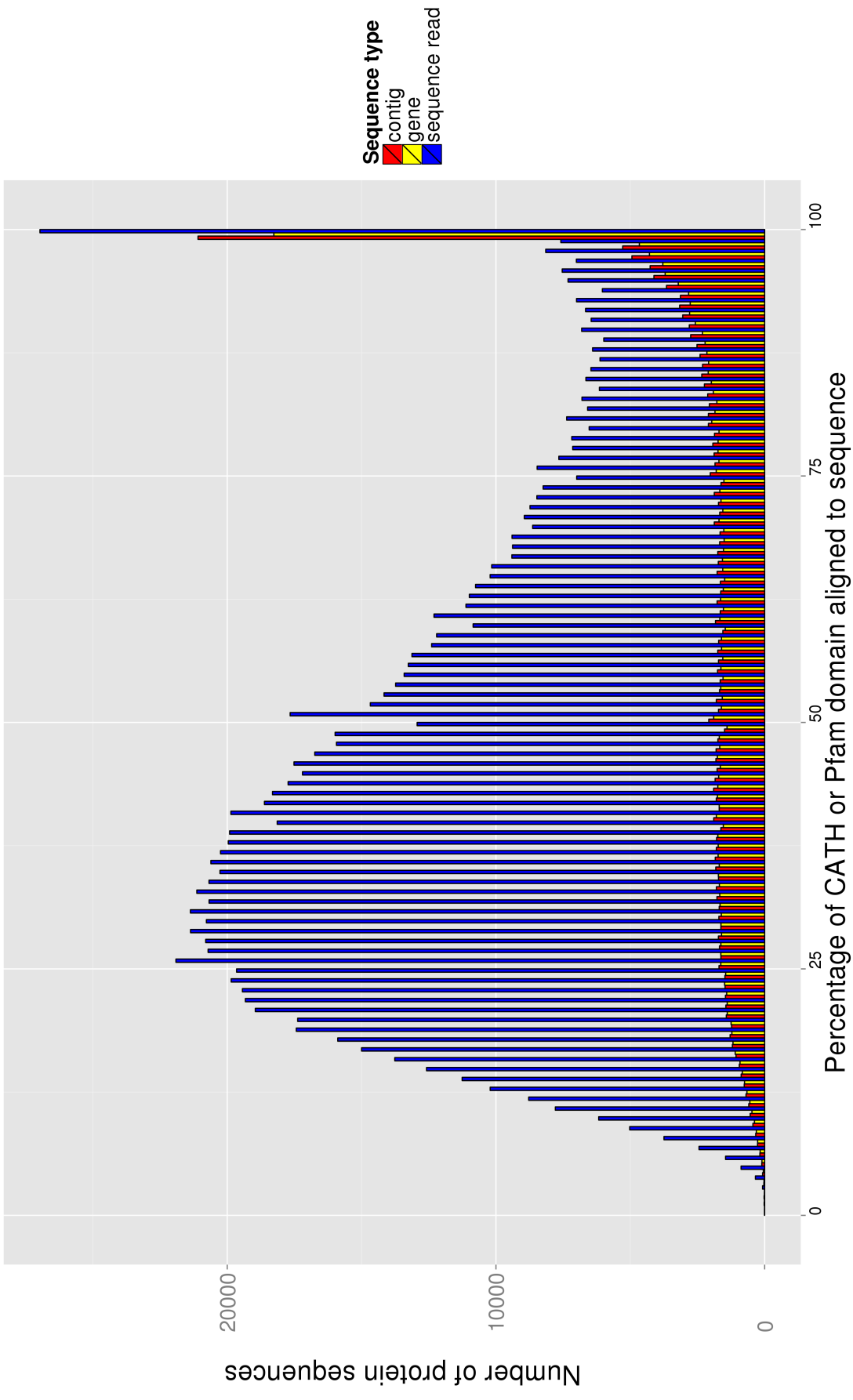


Figure 4.6: The percentage alignment coverage of the protein sequences (either contigs, genes, or sequence reads) against a matched CATH or Pfam domain HMM.

Figure 4.7 compares the number of functional families assigned to three bacteria genomes in Gene3D with the number of functional families assigned to the sequence reads from the oral metagenome data. The three bacterial genomes have around 3-4,000 different functional families and the oral metagenomes have approximately twice as many functional family assignments.

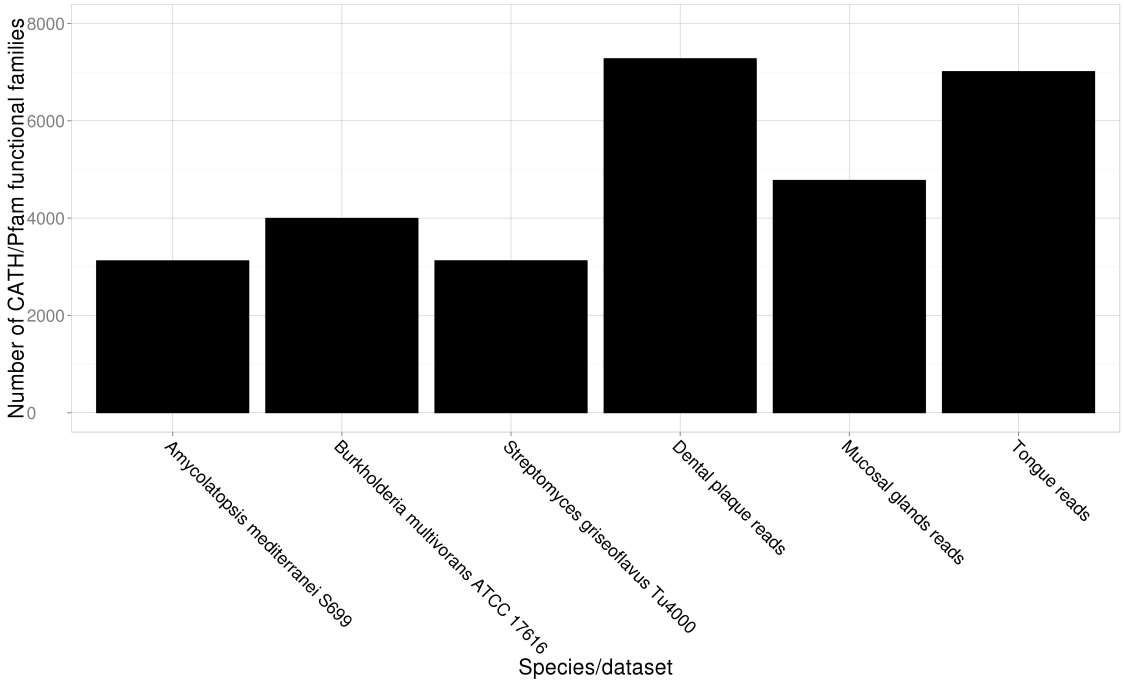


Figure 4.7: The number of CATH/Pfam functional families identified in protein domain sequences from three bacteria in Gene3D and in sequence reads from the three oral metagenomes.

4.3.5.2 Providing an overview of the functional family assignments by categorising them using broad GO term information

To overcome the issue of low functional family coverage for sequence reads data, the coverage was increased by combining the functional families assigned to gene sequences with functional families identified in oral bacterial strains. The oral microbiomes were identified by determining the bacterial strains present (see Methods Section 4.2.4) and then extracting the sequences associated with those strains in Gene3D. These sequences were then mapped to their corresponding functional families using the in-house Gene3D mapping.

Through the combination of gene- and species-derived FunFams the sequence data were mapped to 18,149, 12,765, and 22,645 FunFams for the dental plaque, mucosal glands, and the tongue data, respectively. This increased the functional family coverage from $\sim$ 1-3% to $\sim$ 11-20%.

Figure 4.8 shows the combined set of FunFams grouped into broad GO categories from the molecular function ontology. Similar numbers of functional family assignments are categorised into the same broad GO terms for all three oral environments: dental plaque, mucosal glands, and tongue. The most abundant GO terms are 'binding' and 'catalytic activity', which each categorise thousands of functional families. The term binding is very general and simply refers to any interaction made by a protein with another molecule at one or more specific sites (Binns *et al.*, 2009). The bacterial communities in each of these three environments will need to express proteins that allow them to recognise and attach to a host surface for example, and work together with other members of the community to form a biofilm. Bacteria will also bind nutrients before degrading them using various different types of enzymes including peptidases and glycoside hydrolases. The latter are just a few examples of the types of catalytic activity that are likely be occurring within each environment.

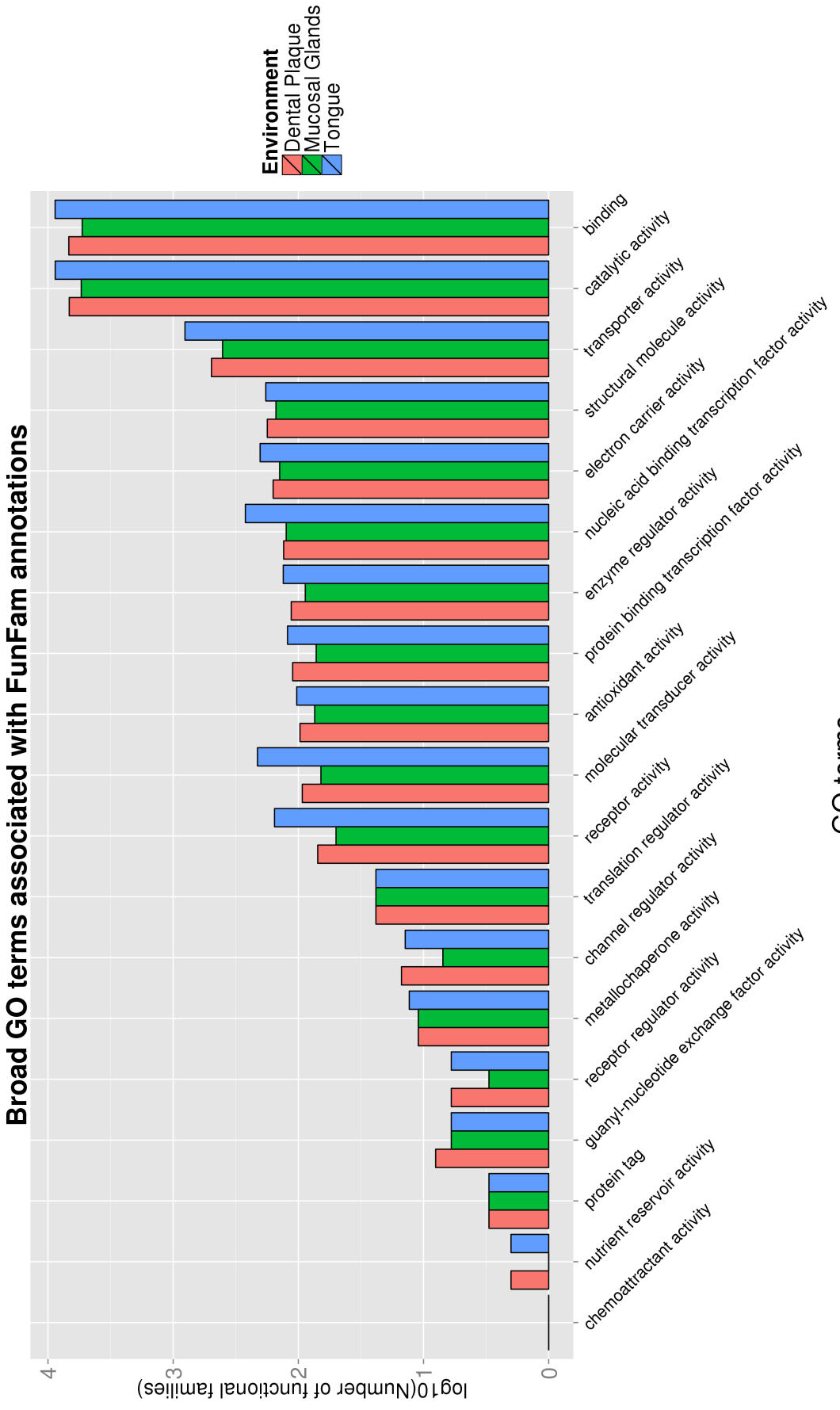


Figure 4.8: The number of FunFams categorised into broad GO terms from the Molecular Function Ontology.

4.3.6 Comparing the functional profiles of microbiomes

To explore any significant differences in the functional repertoires of bacteria in a healthy oral and a healthy gut environments, the functional annotations associated with the tongue and gut metagenomes were compared. The tongue and gut metagenomes were also compared against a general bacterial background (consisting of 1,438 genomes that map to 81,989 CATH and Pfam FunFams) to explore how the functional abilities of the microbiomes differ from that of typical bacteria.

Comparing FunFam abundance in the tongue metagenome against the bacterial background Comparing FunFam abundance between the tongue microbiome and the bacterial background, 1860 FunFams (892 CATH and 968 Pfam) were enriched in the tongue microbiome, and 1077 FunFams (642 CATH and 435 Pfam) were enriched in the bacterial background. Table 4.10 shows the top ten most significantly abundant FunFams in the tongue. There is a clear enrichment in bacterial proteins involved in cell wall formation (Penicillin binding proteins), cell wall binding (Putative cell wall binding repeat, Coiled stalk adhesin, PASTA domain, YSIRK type signal peptide), and in the transport of molecules in and out of the cell (Outer membrane porin). These abundances could be a result of the harsh oral environment where bacteria are frequently removed through swallowing.

CATH/Pfam FunFam	**FunFam Name**	**P-value**
2.40.160.10_9688	Outer membrane porin	0
3.40.710.10_4277	Penicillin binding protein 2B	0
3.40.710.10_4880	Penicillin-binding protein 2x (EC 2.4.1.129)	0
PF01473_6147	Putative cell wall binding repeat	0
PF03793_3989	PASTA domain	0
PF04650_542	YSIRK type signal peptide	0
PF05662_2126	Coiled stalk of trimeric autotransporter adhesin	0
PF07673_202	DUF1602	0
PF08794_68	Lipoprotein GNA1870 C terminal like	0
PF00521_2877	DNA gyrase/topoisomerase IV, subunit A	2.21E-306

Table 4.10: The top ten most significantly abundant CATH and Pfam FunFams in the tongue microbiome, in comparison with the bacterial background.

Comparing FunFam abundance in the gut metagenome against the bacterial background Comparing FunFam abundance between the gut microbiome and the bacterial background, 3630 (1688 CATH and 1942 Pfam) FunFams were enriched in the gut microbiome, and 4863 (2485 CATH and 2378 Pfam) FunFams were enriched in the bacterial background. Table 4.11 shows the top ten most significantly abundant FunFams in the gut. These functional annotations are very different to those found to be enriched in the tongue metagenome and describe a number of enzymes involved in the reduction of nitrites.

CATH/Pfam FunFam	**FunFam Name**	**P-value**
1.20.1450.10_1	Ammonia monooxygenase subunit A-1	0
1.20.1450.10_11	Particulate methane monooxygenase beta subunit	0
2.130.10.10_102414	Nitrous oxide reductase	0
2.140.10.20_2738	Nitrite reductase (EC 1.7.2.1) (Cytochrome cd1)	0
2.40.30.10_37387	Elongation factor Tu	0
2.60.40.420_24014	Copper-containing nitrite reductase NirK (EC 1.7.2.1)	0
2.60.40.420_24056	Nitrous oxide reductase	0
3.20.20.110_826	Ribulose bisphosphate carboxylase large chain 1	0
3.30.1510.10_800	C-1-tetrahydrofolate synthase, cytoplasmic	0
3.30.260.10_2157	60 kDa heat shock protein, mitochondrial	0

Table 4.11: The top ten most significantly abundant CATH and Pfam FunFam in the gut microbiome, in comparison with the bacterial background.

Comparing the tongue metagenome against the gut metagenome In the comparison of FunFam abundance between the tongue and gut microbiomes, 1300 FunFams (749 CATH and 551 Pfam) were enriched in the tongue microbiome, and 733 FunFams (248 CATH and 485 Pfam) were enriched in the gut microbiome.

Tables 4.12 and 4.13 list the top ten most significantly abundant FunFams in the tongue and gut microbiome, respectively. The results obtained in comparing the metagenomes directly are similar as to when comparing the metagenomes against the bacterial background. One noticeable difference is the presence of the 'Glycosyl hydrolase (GH) family 2, sugar binding domain', which suggests an enrichment of carbohydrate degradation in the gut metagenome. This is not a surprising result as other groups have also reported an enrichment in carbohydrate degradation enzymes.

CATH/Pfam FunFam	**FunFam Name**	**P-value**
2.40.160.10_9688	Outer membrane porin	0
PF07673_202	DUF1602 (clan)	0
PF01473_6147	Putative cell wall binding repeat	1.148025E-222
PF05662_2126	Coiled stalk of trimeric autotransporter adhesin	5.900531E-218
PF13900_908	Putative binding domain	1.879078E-196
3.40.710.10_4880	Penicillin-binding protein 2x (EC 2.4.1.129)	7.441453E-181
PF08794_68	Lipoprotein GNA1870 C terminal like	6.429207E-157
2.60.40.1380_1148	LPXTG cell wall surface protein	5.459921E-150
PF04650_542	YSIRK type signal peptide	4.759559E-146
3.40.710.10_4277	Penicillin binding protein 2B	6.563816E-144

Table 4.12: The top ten most significantly abundant CATH and Pfam FunFams in the tongue microbiome, in comparison with the gut microbiome.

CATH/Pfam FunFam	**FunFam Name**	**P-value**
2.140.10.20_2738	Nitrite reductase (EC 1.7.2.1)	1.320525E-219
3.40.50.300_26153	Nitrogenase iron protein 1 (EC 1.18.6.1)	1.523918E-144
PF02837_3829	Glycosyl hydrolases family 2, sugar binding domain	8.719111E-140
3.40.50.300_27595	Nitrogenase iron protein (EC 1.18.6.1)	1.579457E-130
PF14310_3137	Fibronectin type III-like domain	7.294723E-126
2.130.10.10_102414	Nitrous oxide reductase	4.692666E-120
PF06541_1134	DUF1113	4.110635E-115
3.20.20.110_826	Ribulose-bisphosphate carboxylase (EC 4.1.1.39)	2.096357E-100
3.40.50.740_11858	Nitrate reductase 2 (NRZ), alpha subunit	3.489883E-092
PF14198_368	Transposon-encoded protein TnpV	5.245022E-092

Table 4.13: The top ten most significantly abundant CATH and Pfam FunFams in the gut microbiome, in comparison with the tongue microbiome.

As in previous comparisons against the general bacterial background, we also see an enrichment in adhesion proteins in the tongue microbiome and an enrichment of two nitrogenase iron proteins in the gut microbiome, which are components of the nitrogenase enzyme system (EC 1.18.6.1). This system catalyses the reduction of dinitrogen to ammonia.

4.3.6.1 Identifying significantly enriched metabolic genes and pathways in the tongue and gut data

In order to define an enriched set of metabolic genes in the tongue and gut metagenomes, the two data sets of sequence reads were annotated using the KO database through the MG-RAST server. This resulted in 1117 unique KO annotations for the tongue data, and 1931 KO annotations for the gut data.

Using the Fisher exact test described in Methods (see page 186), a total of 122 and 48 KO terms were found to be significantly more abundant in the tongue metagenome and in the gut metagenome, respectively. These enriched terms were then used to explore whether their metabolic pathways were also enriched in the metagenome using Equation 4.2 in Section 4.2.7.2.

The 15 most enriched KEGG pathways are shown in Tables 4.14 and 4.15. It should be noted that while different KO terms are enriched between the two data sets, different KO terms may map to the same KEGG pathway.

Pathway ID	**Pathway name**	**PES**	**KOs enriched**	**KOs in pathway**	**List of enriched KOs**
ko00240	Pyrimidine metabolism	10.14	11	168	K00526 K00762 K00857 K00940 K01489 K01493 K01520 K02825 K03046 K03763 K09903
ko00790	Folate biosynthesis	8.12	5	35	K00950 K01633 K01737 K03637 K03639
ko00230	Purine metabolism	4.55	11	254	K00526 K00759 K00760 K00939 K00940 K01515 K01588 K03046 K03763 K03816 K11175
ko01230	Biosynthesis of amino acids	3.55	9	225	K00640 K00891 K01653 K01658 K01750 K01808 K01817 K01958 K03786
ko00983	Drug metabolism - other enzymes	3.41	3	22	K00760 K00857 K01489
ko00620	Pyruvate metabolism	2.76	5	87	K00162 K01678 K01759 K01958 K02160
ko00910	Nitrogen metabolism	2.67	3	56	K00284 K00370 K00373
ko00400	Phenylalanine, tyrosine and tryptophan biosynthesis	2.32	4	72	K00891 K01658 K01817 K03786
ko00720	Carbon fixation pathways in prokaryotes	2.26	5	97	K01678 K01848 K01958 K02160 K03518
ko00770	Pantothenate and CoA biosynthesis	2.04	3	34	K00867 K00997 K01653
ko01200	Carbon metabolism	1.99	8	308	K00162 K00640 K01678 K01808 K01848 K01958 K02160 K03518
ko00540	Lipopolysaccharide biosynthesis	1.86	3	34	K02517 K02843 K03270
ko00903	Limonene and pinene degradation	1.63	2	19	K00680 K01692
ko00020	Citrate cycle (TCA cycle)	1.32	3	54	K00162 K01678 K01958
ko00051	Fructose and mannose metabolism	1.25	3	90	K01808 K02794 K02795
ko00270	Cysteine and methionine metabolism	1.25	3	81	K00558 K00640 K00641
ko00633	Nitrotoluene degradation	1.24	2	17	K03518 K06281
ko00564	Glycerophospholipid metabolism	1.18	4	99	K00901 K00980 K00995 K08591
ko00640	Propanoate metabolism	1.17	3	76	K01692 K01848 K02160
ko00052	Galactose metabolism	1.09	3	72	K01819 K02773 K02774

Table 4.14: The top 20 enriched metabolic pathways (out of 86) in the tongue microbiome. PES is the pathway enrichment score.

Pathway ID	Pathway name	PES	KOs enriched	KOs in pathway	List of enriched KOs
ko01210	2-Oxocarboxylic acid metabolism	4.61	7	74	K00031 K00053 K00928 K01647 K01681 K01687 K01703
ko01230	Biosynthesis of amino acids	4.44	12	225	K00031 K00053 K00265 K00928 K01586 K01647 K01681 K01687 K01703 K01755 K01915 K01940
ko00250	Alanine, aspartate and glutamate metabolism	4.01	6	65	K00262 K00265 K01755 K01915 K01940 K01955
ko00290	Valine, leucine and isoleucine biosynthesis	3.51	3	17	K00053 K01687 K01703
ko00052	Galactose metabolism	2.48	5	72	K00965 K01187 K01190 K07407 K12308
ko00020	Citrate cycle (TCA cycle)	2.02	4	54	K00031 K01610 K01647 K01681
ko00630	Glyoxylate and dicarboxylate metabolism	1.73	4	74	K01647 K01681 K01915 K01966
ko00620	Pyruvate metabolism	1.67	5	87	K00656 K01572 K01595 K01610 K04072
ko00500	Starch and sucrose metabolism	1.6	4	96	K00688 K00705 K01187 K05349
ko00511	Other glycan degradation	1.31	2	18	K01190 K01206
ko00720	Carbon fixation pathways in prokaryotes	1.16	4	97	K00031 K01595 K01681 K01966
ko00910	Nitrogen metabolism	1.15	3	56	K00262 K00265 K01915
ko00330	Arginine and proline metabolism	0.93	4	133	K00262 K01755 K01915 K01940
ko00071	Fatty acid degradation	0.91	2	50	K01897 K04072
ko00770	Pantothenate and CoA biosynthesis	0.83	2	34	K00053 K01687
ko01212	Fatty acid metabolism	0.78	2	69	K01897 K11533
ko00600	Sphingolipid metabolism	0.77	2	41	K01190 K07407
ko01200	Carbon metabolism	0.76	6	308	K00031 K01595 K01610 K01647 K01681 K01966
ko00603	Glycosphingolipid biosynthesis - globo series	0.65	1	14	K07407
ko00460	Cyanoamino acid metabolism	0.61	1	31	K05349

Table 4.15: The top 20 enriched metabolic pathways (out of 64) in the gut microbiome. PES is the pathway enrichment score.

There are few studies comparing the functional annotations of the oral and gut metagenomes. Belda-Ferre *et al.* (2011) compared their own oral metagenome data with the Kurokawa *et al.* (2007) gut data set and concluded that metabolic genes involved in sugar uptake and assimilation, adhesion proteins, and prophage genes were enriched in the gut microbiome. Whereas in the oral microbiome, enriched metabolic genes were involved in oxidative and osmotic stress, or in iron scavenging.

In agreement with Belda-Ferre *et al.* (2011), a number of enriched metabolic pathways involved in carbohydrate metabolism have been found in the gut metagenome used in this work. We have found enriched metabolic enzymes in the gut (relative to the tongue) that map to 11 carbohydrate metabolism KEGG pathways: Glycolysis / Gluconeogenesis, Citrate cycle (TCA cycle), Galactose metabolism, Starch and sucrose metabolism, Amino sugar and nucleotide sugar metabolism, Pyruvate metabolism, Glyoxylate and dicarboxylate metabolism, Propanoate metabolism, Butanoate meta-bolism, C5-Branched dibasic acid metabolism, and the Pentose and glucuronate interconversions pathway.

One of the carbohydrate metabolism pathways with the highest pathway enrichment score is the Starch and sucrose metabolism KEGG pathway. Large amounts of undigested plant polysaccharides (e.g. cellulose, xylan, and pectin) and partially digested starch reach the gut which cannot be digested by host enzymes. Members of the gut microbiome however are able to break down such carbohydrates, for example members of the Bacteroidetes phyla, through the use of enzymes including glycoside hydrolases and polysaccharide lyases, which form part of the Starch and sucrose metabolism pathway. The gut microbiota are also known to produce short chain fatty acids (SCFAs) from these complex carbohydrates, which could explain the higher enrichment scores in lipid metabolism pathways in the gut (den Besten *et al.*, 2013; Devaraj *et al.*, 2013).

Enriched KO terms/genes involved in carbohydrate metabolism have also been found in the tongue metagenome. Once carbohydrates have been degraded, bacterial systems are required to uptake the products. There are enriched KO genes in the

tongue metagenome which map to the Phosphotransferase system (PTS) pathway. This system is widely used by bacteria to uptake around 20 different carbohydrates such as hexoses, hexitols, and disaccharides (Kotrba *et al.*, 2001;Kanehisa *et al.*, 2014).

Enzymes within the denitrification system were found to be enriched in the gut metagenome using the FunFam annotations. The KO gene analysis reports that there are enriched genes in the gut as well as the tongue that map to the Nitrogen metabolism KEGG pathway, which the denitrification system is part of. Despite the enriched genes from the gut and the tongue mapping to the same KEGG pathway we can identify different parts of this pathway that are being performed in these two different environments. Figure 4.9 highlights the enriched enzymes reported from the FunFam and KO enrichment studies within the nitrogen metabolism pathway.

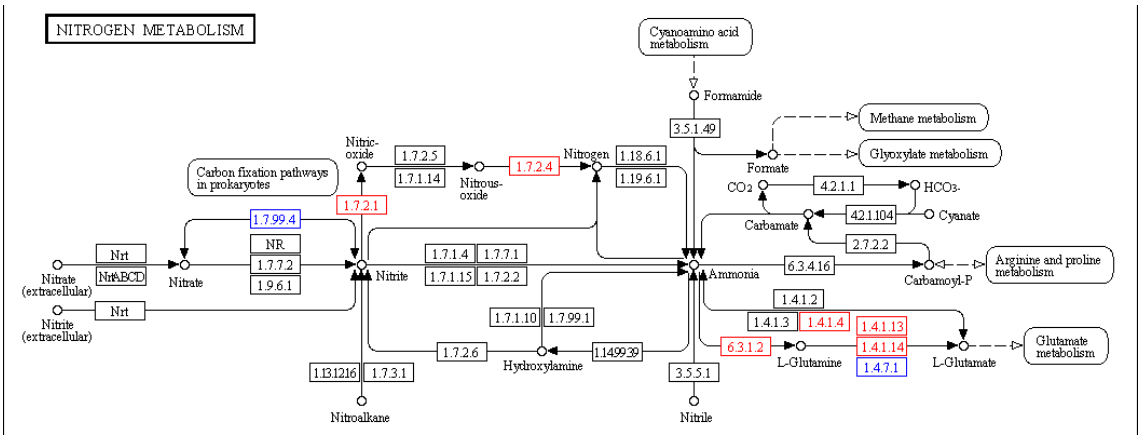


Figure 4.9: The enriched FunFam and KO terms mapped onto the nitrogen metabolism KEGG pathway. Enriched terms in the tongue metagenome are highlighted in blue and in the gut metagenome in red.

In the tongue metagenome, the three enriched KO terms in this pathway relate to: 1) nitrate reductase (EC 1.7.99.4), which reduces nitrate to nitrite, 2) glutamate synthase (EC 1.4.7.1), which reduces L-glutamate to L-glutamine (which is then reduced to ammonia by a different enzyme), and 3) nitrate reductase alpha subunit (EC 1.7.99.4), which is part of nitrate reductase (see Figure 4.9). The three enriched KO terms in the gut metagenome relate to: 1) glutamate synthase (EC 1.4.1.13, EC 1.4.1.14), 2) glutamate dehydrogenase (NADP+) (EC 1.4.1.4), and 3) glutamine

synthase (EC 6.3.1.2). All these enzymes are highlighted in Figure 4.9.

A specific part of the nitrogen metabolism pathway, the denitrification system (see Figure 4.10), has previously been shown to be beneficial to human health. The ingested nitrate is reduced to nitrite by the oral bacteria. The nitrite is taken up into the host's bloodstream in the gut and converted into nitric oxide. Nitric oxide is essential for healthy blood vessels as it helps to keep them supple, which helps maintain a low blood pressure (Wade, 2013). It is reassuring to observe that our results reflect this process by showing that the enzymes catalysing the reaction of nitrate to nitrite is enriched in the tongue metagenome and showing that the enzyme catalysing nitrite to nitric oxide is enriched in the gut metagenome (see Figure 4.10).

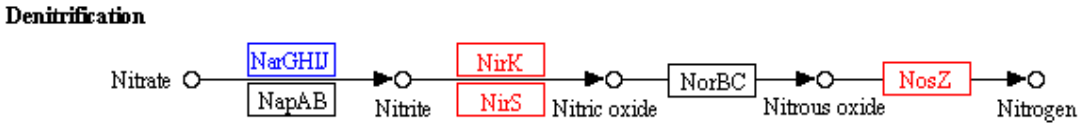


Figure 4.10: The enriched FunFam and KO terms mapped onto the denitrification system in KEGG. Enriched terms in the tongue metagenome are highlighted in blue and in the gut metagenome in red.

4.4 Conclusions and future work

In this chapter, metagenome data from different areas of the oral cavity were processed and characterised in terms of bacterial species present and protein function annotations.

The bacterial phyla classified in the three oral metagenome data are consistent with reports from other studies. Dewhirst *et al.* (2010) found that the microbiome of the mouth is dominated by the bacterial phyla (in descending order) Firmicutes, Bacteroidetes, Proteobacteria, Actinobacteria, Spirochaetes, and Fusobacteria. This work reports a similar pattern in the three oral metagenomes studied with minor variations. For example the most predominant phylum in the dental plaque and tongue metagenomes is the Proteobacteria rather than the Firmicutes. In addition, the Spirochaetes phylum is only reported in the dental plaque and tongue metagenomes and with quite low counts of only 87 and 1 sequence read(s), respectively. Therefore there may not have been enough sequence read coverage of these bacterial genomes in the mucosal glands data set, resulting in an apparent absence of this phylum.

Functional annotations have been compared between the healthy tongue metagenome, a publicly-available healthy gut metagenome, and a general bacterial background to understand firstly, what is enriched within each metagenome, and secondly, to investigate what is enriched between the two metagenomes.

Functional family (FunFam) and KEGG Orthology (KO) terms were used in this analysis to explore whether certain functions are enriched within a particular metagenome environment. Differences between the tongue and the gut metagenomes were first analysed using the FunFams. The most enriched FunFams in the tongue were involved in cell wall formation and binding, whereas in the gut the most enriched FunFams, compared to the tongue, were mainly involved in the denitrification process.

Enriched metabolic genes in the tongue and gut metagenomes both map to metabolic pathways involved in the fermentation of amino acids into short chain

fatty acids. Fusobacterium have aminopeptidase activity, which have been found in small quantities in the tongue metagenome, and they have been reported to produce short chain fatty acids (SCFA) following fermentation (Wade, 2013). As previously mentioned, the gut microbiota are able to produce SCFA from complex carbohydrates such as starch.

In this analysis we have also searched for enriched genes via KEGG Orthology gene functional annotations and we have mapped these enriched genes to metabolic pathways. While this analysis is useful for identifying and comparing enriched KO genes in either the tongue or the gut microbiomes, it is rather difficult to compare the enrichment of metabolic pathways between metagenomes. This is because the different enriched genes in each metagenome can be mapped to the same metabolic pathway. Therefore a given metabolic pathway can potentially be enriched in both metagenomes and further analysis is needed to determine which parts of the pathway are enriched in one or the other metagenome, as was done with the nitrogen metabolism pathway, above. The pathway enrichment score has therefore simply been used as an indication of how likely it is that a given pathway is present in that metagenome.

A number of other complications are associated with the study of metagenome data. In this analysis, the DNA extraction process for all three oral metagenome data sets was the same, however it is not known whether the process is exactly the same as the one used to generate the gut dataset as Yatsunenko *et al.* (2012) simply states that a common DNA extraction protocol was used. The oral and gut data were sequenced in different laboratories, which may have also had an impact. However, although the biodiversity of the compared tongue and gut data are not identical, they both contain hundreds of species. The complexity of the two samples is therefore not too dissimilar. It is also reassuring to observe similar enriched functional terms in the tongue and gut in comparison to other studies such as Belda-Ferre *et al.* (2011).

Another confounding factor is the issue of unequal genome coverage for all bacteria in the data set. As mentioned previously, the Spirochaetes phylum was only

detected in the dental plaque metagenome, and in small numbers. It could be that we have not been able to detect a number of bacterial genomes due to low coverage and it would be interesting to compile a list of all previously characterised oral bacterial species in these metagenomes, to compare our results against. However this would not account for uncharacterised species that may be under-represented by genomic coverage.

If a dataset has uneven sequence coverage of a metagenome, the length and quality of the subsequent contigs assembled and the genes predicted from this will be affected. If one looks at the number of sequence reads assembled for the three oral metagenome data, we can see that only 39-67% of the data was used. This may be a result of missing sequence reads due to low genomic coverage or the removal of reads during the quality assessment. Clearly, this reduces the number of gene predictions that can be made, which in turn affects the analysis. This is reflected in Figure 4.11 where we show that the number of different functional families identified from predicted metagenome genes is similar to the number of functional families identified in the protein domain sequences of a single bacterial genome, whereas we would expect a larger number of functional families to be detected in a metagenome. As in this work, the literature has also reported that there are many hundreds of bacterial genomes in the oral metagenomes, and therefore there should be many times more genes predicted within the metagenome data and many more functional families identified.

We are now testing different assembly methods to find out whether we can assemble larger percentages of the data. While it is difficult to assess the quality of these *de novo* sequence assemblies, as we do not have reference genomes for all the bacteria in the oral and gut environments, we could also try mapping the data onto reference genome sequences. For example, we could download all of the characterised bacterial genomes from the HMP for these oral environments and map the sequence reads against them to produce contigs. This would allow us to directly assess their sequence quality.

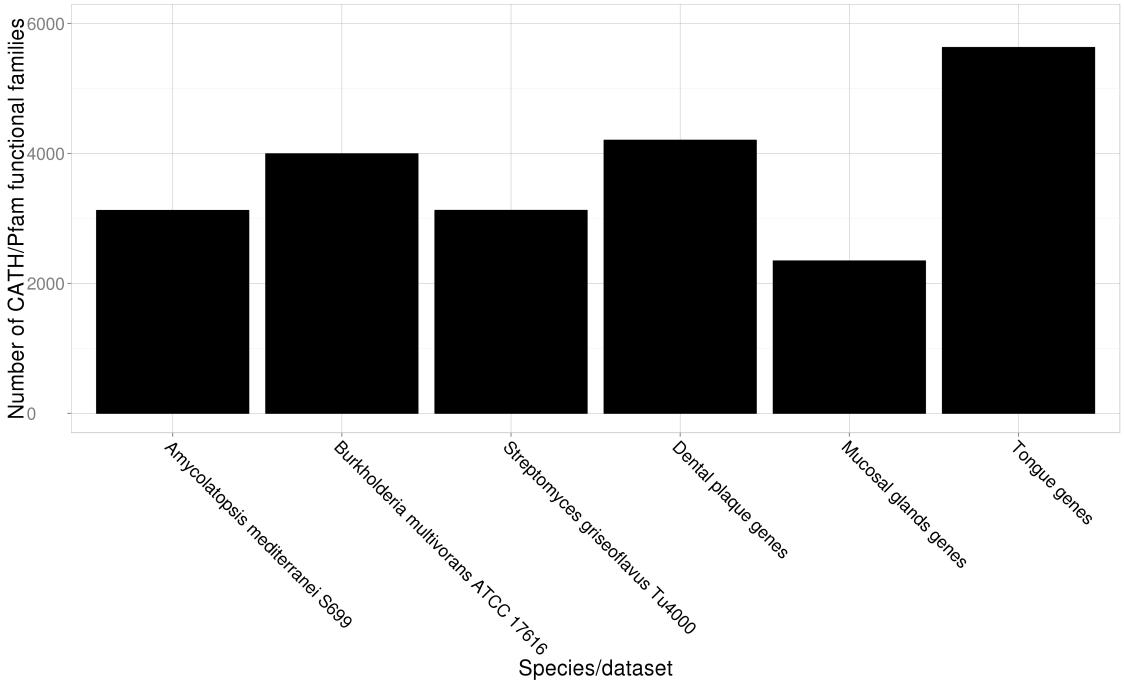


Figure 4.11: The number of CATH/Pfam functional families identified in protein domain sequences from three bacteria in Gene3D and in gene predictions from the three oral metagenomes.

While the metagenome data used for this work was sequenced using 454 technology, problems can arise in metagenomic analysis from the comparison of data sets that have used different sequencing technologies. In fact, studies have been published that discourage the comparison of datasets that have used different sequencing technologies or environmental sampling protocols (Mitra *et al.*, 2009). This is due to the fact that statistically different results obtained from such comparative analysis may not reflect true differences between the data but instead report false positives due to these experimental differences.

Approximately 40% of sequence reads from the oral metagenomes could not be classified into a CATH (or a Pfam) superfamily. These sequences are referred to as novel sequences due to their lack of annotation and are classified here as not belonging to either a superfamily or a functional family. It would be interesting to look at the number of sequence reads that hit a superfamily but that do not hit a functional family within that superfamily. These sequences would represent functionally diverse superfamily members. A similar analysis was carried out when

looking for novel sequences in three large gut metagenomic datasets. Ellrott *et al.* (2010) identified $\sim$ 1980 novel protein families (i.e. homologous sequences that were clustered based upon whole-protein sequence similarities) from several human gut metagenomic studies by scanning sequences against Pfam-B families and identifying sequences that did not match any Pfam A or B family. 1800 were automatically clustered and the remaining 180 were manually curated. These were then added to the Pfam database (Godzik, 2011). In our analysis we would be using oral metagenome data instead and would incorporate our findings into the CATH-Gene3D resource.

Finally, as metagenome data provides genomic content that is not necessarily expressed and translated into a protein product, it would be interesting to apply a metatranscriptomics approach to the data to look at the transcripts produced. These transcripts are typically sequenced using Illumina sequencing technology, which has a lower sequence error rate than the 454 technology used in this work (discussed in Section 4.1.3) and a higher sequence read coverage. This would help to improve sequence assembly quality, which should result in more protein domains and genes being identified and annotated.

Chapter 5

Conclusions

Functional diversity across CATH superfamilies has been characterised and quantified to understand how function evolves. Different classifications of functional families, which define different functional groups within CATH superfamilies, were first assessed to determine the most functionally coherent classification of families. Experimentally-determined functional site residue information on binding and catalytic residues was then used to characterise the functional sites used by different relatives across functionally diverse superfamilies.

This work shows, on a much larger scale than previous studies, that functionally diverse relatives tend to have different protein partners and therefore use different protein-protein interface binding sites. We have also shown that despite these differences, there is overlap between regions of these different large interfaces, and this suggested a preference for a particular interface site or surface in many of the superfamilies studied. It would be interesting to analyse further the reasons for this preference in different superfamilies and whether it is a result of a particular structural characteristic of the surface eg planarity or a constraint imposed by a cofactor.

We then use functional families to look more specifically at how function (defined by EC) evolves within enzyme domain superfamilies by examining the relationship between changes in catalytic machinery and changes in reaction mechanism. Catalytic machineries were studied in over one hundred CATH enzyme superfamilies and found to exhibit considerable variation across the majority of the superfamilies studied. Again, this phenomenon has not been previously reported on such a large scale.

When examining whether a change in catalytic machinery is accompanied by a change in reaction mechanism, perhaps the most surprising result was that nearly one-quarter of the functional families compared used different catalytic machineries

to perform the same reaction mechanisms, measured by bond change similarity. While a large proportion of these functional families were associated with a change in substrate, i.e. a change at the fourth EC hierarchical level, a significant proportion of functional families used the same substrate. It would be interesting to perform more detailed phylogenetic studies of these cases to examine whether they are examples of true functional convergence or whether they are due to divergence of one relative, from a common function, followed by convergence back to that function.

In the second work chapter, we also examined the preference for catalytic residues to be present in loop regions. Beta and alpha-beta class superfamilies were shown to have statistically significantly more catalytic residues in loop regions in comparison with alpha class superfamilies. We also examined whether TIM barrel and Rossmann fold superfamilies are more likely to have their catalytic residues in loops. Tawfik and others have suggested that this is the case and that this contributed to their incredible functional diversity as the loops are detached from the beta-sandwich or beta-barrel scaffold so that any mutations that promote a new function are less likely to destabilise the structure. We are currently analysing catalytic residue preference within all CATH superfamilies that have a beta-sandwich or beta-barrel topology. Many superfamilies adopting these structural scaffolds are well-known to support completely different enzymatic functions and therefore different catalytic machineries, and many feature loops that are clustered around their active sites.

In the final work chapter, we presented the functional families as a new and powerful method for annotating metagenome data. Functional families were used to characterise the functions performed by the bacterial communities inside the oral and gut metagenomes, the two most bacterially species-diverse environments in the human body. The results illustrated the different ways in which the bacteria have adapted to their host in order to survive. For example, we found an enrichment of adhesion proteins in the tongue metagenome and enzymes involved in the denitrification system in the gut metagenome. This new functional family annotation protocol is a valuable tool in the field of comparative metagenomics and can help in

understanding how certain bacteria adapt to survive within the human body as commensal and symbiotic species. We are currently testing new metagenome sequence read assembly methods to assemble as much of the sequence data as possible, which is expected to lead to an improvement in the number of gene predictions and the number of functional families that can be identified.

Appendix A

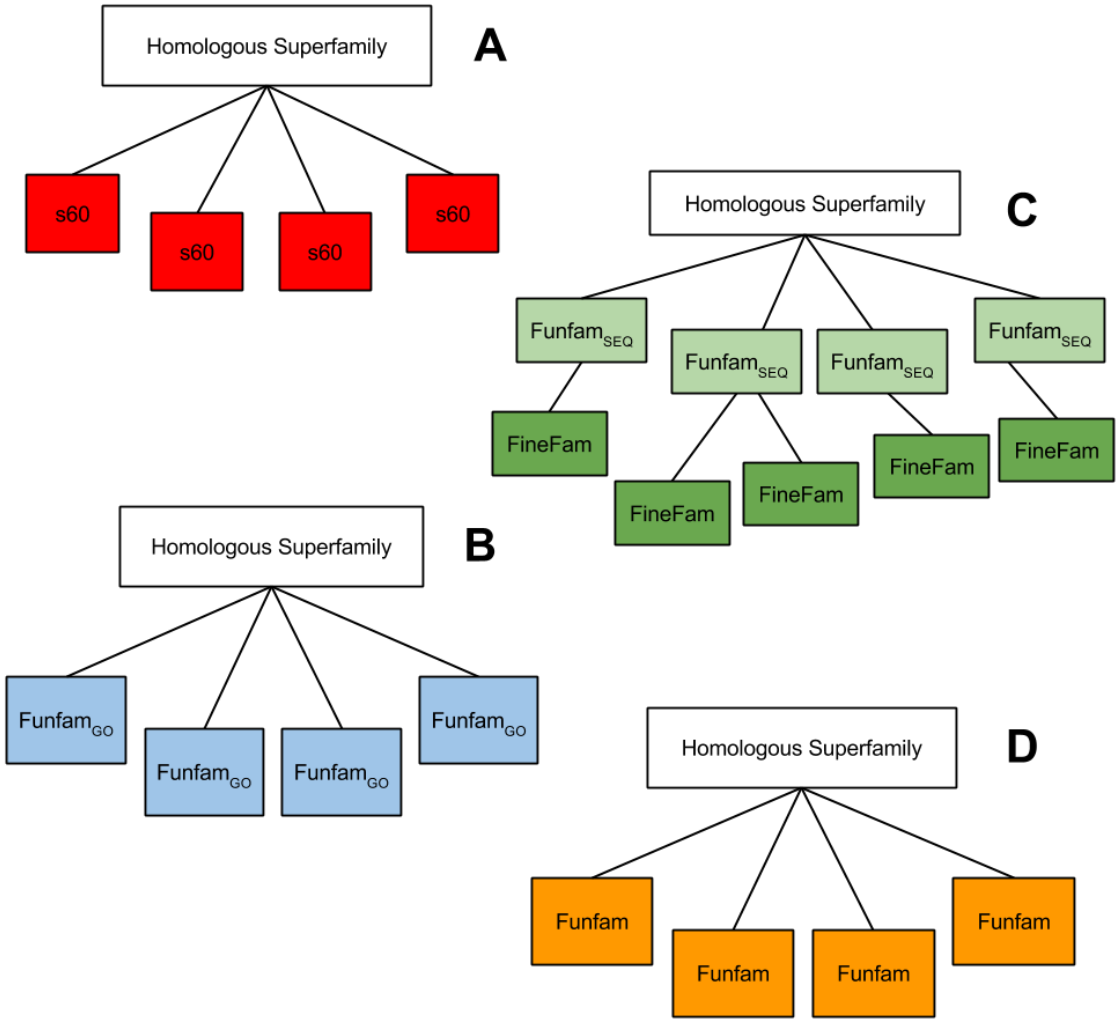


Figure A.1: The subclassification of CATH homologous superfamilies into the different types of functional families studied: sequences clustered at 60% sequence identity, referred to as s60 clusters (A), FunFam$_{GO}$s generated using the DFX method (B), FunFam$_{SEQ}$s generated using the Funfamer method (C), and FunFams generated using the FunFHMMER method (D).

References

Abhiman, S. and Sonnhammer, E. L. (September 2005a). Large-scale prediction of function shift in protein families with a focus on enzymatic function., *Proteins*, **60**.4, 758–768. 41, 55, 56

Abhiman, S. and Sonnhammer, E. L. L. (January 2005b). FunShift: a database of function shift analysis on protein subfamilies, *Nucleic Acids Research*, **33**.suppl 1, D197–D200. 55

Addou, S., Rentzsch, R., Lee, D. and Orengo, C. A. (March 2009). Domain-Based and Family-Specific Sequence Identity Thresholds Increase the Levels of Reliable Protein Function Transfer, *Journal of Molecular Biology*, **387**.2, 416–430. 38

Akiva, E., Brown, S., Almonacid, D. E., Alan, Custer, A. F., Hicks, M. A., Huang, C. C., Lauck, F., Mashiyama, S. T., Meng, E. C., Mischel, D., Morris, J. H., Ojha, S., Schnoes, A. M., Stryke, D., Yunes, J. M., Ferrin, T. E., Holliday, G. L. and Babbitt, P. C. (January 2014). The StructureFunction Linkage Database, *Nucleic Acids Research*, **42**.D1, D521–D530. 48

Almonacid, D. E., Yera, E. R., Mitchell, J. B. O. and Babbitt, P. C. (March 2010). Quantitative Comparison of Catalytic Mechanisms and Overall Reactions in Convergently Evolved Enzymes: Implications for Classification of Enzyme Function, *PLoS Computational Biology*, **6**.3, e1000700+. 110

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (October 1990). Basic local alignment search tool., *Journal of Molecular Biology*, **215**.3, 403–410. 38

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (September 1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs., *Nucleic Acids Research*, **25**.17, 3389–3402. 39, 50

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E. and Murzin, A. G. (January 2014). SCOP2 prototype: a new approach to protein structure mining, *Nucleic Acids Research*, **42**.D1, D310–D314. 23, 26

Anton, B. P., Chang, Y.-C., Brown, P., Choi, H.-P., Faller, L. L., Guleria, J., Hu, Z., Klitgord, N., Levy-Moonshine, A., Maksad, A., Mazumdar, V., McGettrick, M., Osmani, L.,

Pokrzywa, R., Rachlin, J., Swaminathan, R., Allen, B., Housman, G., Monahan, C., Rochussen, K., Tao, K., Bhagwat, A. S., Brenner, S. E., Columbus, L., de Crécy-Lagard, V., Ferguson, D., Fomenkov, A., Gadda, G., Morgan, R. D., Osterman, A. L., Rodionov, D. A., Rodionova, I. A., Rudd, K. E., Söll, D., Spain, J., Xu, S.-y., Bateman, A., Blumenthal, R. M., Bollinger, J. M., Chang, W.-S., Ferrer, M., Friedberg, I., Galperin, M. Y., Gobeill, J., Haft, D., Hunt, J., Karp, P., Klimke, W., Krebs, C., Macelis, D., Madupu, R., Martin, M. J., Miller, J. H., O'Donovan, C., Palsson, B., Ruch, P., Setterdahl, A., Sutton, G., Tate, J., Yakunin, A., Tchigvintsev, D., Plata, G., Hu, J., Greiner, R., Horn, D., Sjölander, K., Salzberg, S. L., Vitkup, D., Letovsky, S., Segrè, D., DeLisi, C., Roberts, R. J., Steffen, M. and Kasif, S. (August 2013). The COMBREX Project: Design, Methodology, and Initial Results, *PLoS Biology*, **11**.8, e1001638+. 48, 49

Aravind, L., Anantharaman, V. and Koonin, E. V. (July 2002). Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA., *Proteins*, **48**.1, 1–14. 57

Arjunan, P., Umland, T., Dyda, F., Swaminathan, S., Furey, W., Sax, M., Farrenkopf, B., Gao, Y., Zhang, D. and Jordan, F. (March 1996). Crystal structure of the thiamin diphosphate-dependent enzyme pyruvate decarboxylase from the yeast Saccharomyces cerevisiae at 2.3 A resolution., *Journal of Molecular Biology*, **256**.3, 590–600. 142

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M. M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H. B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E. G., Wang, J., Guarner, F., Pedersen, O., de Vos, W. M., Brunak, S., Doré, J., MetaHIT Consortium, Antolín, M., Artiguenave, F., Blottiere, H. M., Almeida, M., Brechot, C., Cara, C., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Foerstner, K. U., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Huber, W., van Hylckama-Vlieg, J., Jamet, A., Juste, C., Kaci, G.,

Knol, J., Lakhdari, O., Layec, S., Le Roux, K., Maguin, E., Mérieux, A., Melo Minardi, R., M'rini, C., Muller, J., Oozeer, R., Parkhill, J., Renault, P., Rescigno, M., Sanchez, N., Sunagawa, S., Torrejon, A., Turner, K., Vandemeulebrouck, G., Varela, E., Winogradsky, Y., Zeller, G., Weissenbach, J., Ehrlich, S. D. and Bork, P. (May 2011). Enterotypes of the human gut microbiome., *Nature*, **473**.7346, 174–180. 174, 175

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (May 2000). Gene Ontology: tool for the unification of biology, *Nature Genetics*, **25**.1, 25–29. 29, 30, 49

Ashkenazy, H., Erez, E., Martz, E., Pupko, T. and Ben-Tal, N. (July 2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids, *Nucleic Acids Research*, **38**.suppl 2, W529–W533. 54

Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A., Maudling, N., Mitchell, A. L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A. and Zygouri, C. (January 2003). PRINTS and its automatic supplement, prePRINTS, *Nucleic Acids Research*, **31**.1, 400–402. 27

Attwood, T. K. (September 2002). The PRINTS database: a resource for identification of protein families., *Briefings in Bioinformatics*, **3**.3, 252–263. 40

Babbitt, P. C. and Gerlt, J. A. (December 1997). Understanding enzyme superfamilies. Chemistry As the fundamental determinant in the evolution of new catalytic activities., *Journal of Biological Chemistry*, **272**.49, 30591–30594. 102, 144, 162

Bairoch, A. (January 2000). The ENZYME database in 2000, *Nucleic Acids Research*, **28**.1, 304–305. 28, 31

Barrett, A. J., (1992). *Enzyme Nomenclature: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*. Academic Press, San Diego, California. 28, 30, 31

Bartlett, G. J., Porter, C. T., Borkakoti, N. and Thornton, J. M. (November 2002). Analysis of catalytic residues in enzyme active sites., *Journal of Molecular Biology*, **324**.1, 105–121. 101, 159

Bashton, M. and Chothia, C. (January 2002). The geometry of domain combination in proteins., *Journal of Molecular Biology*, **315**.4, 927–939. 89

Belda-Ferre, P., Alcaraz, L. D., Cabrera-Rubio, R., Romero, H., Simon-Soro, A., Pignatelli, M. and Mira, A. (June 2011). The oral metagenome in health and disease, *The ISME Journal*. 177, 215, 219

Benaglia, T., Chauveau, D., Hunter, D. R. and Young, D. (2009). mixtools: An R package for analyzing finite mixture models, *Journal of Statistical Software*, **32**.6, 1–29. 183

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Sayers, E. W. (January 2013). GenBank, *Nucleic Acids Research*, **41**.D1, D36–D42. 22

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S.,

Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G.-D. D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R. and Smith, A. J. (November 2008). Accurate whole human genome sequencing using reversible terminator chemistry., *Nature*, **456**.7218, 53–59. 167

Berg, J. M., Tymoczko, J. L. and Stryer, L., (2002). *Biochemistry, 5th edition.* W. H. Freeman and Company. 99, 100

Bergthorsson, U., Andersson, D. I. and Roth, J. R. (October 2007). Ohno's dilemma: Evolution of new genes under continuous selection, *Proceedings of the National Academy of Sciences of the United States of America*, **104**.43, 17004–17009. 36

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M. (May 1977). The Protein Data Bank: a computer-based archival file for macromolecular structures., *Journal of Molecular Biology*, **112**.3, 535–542. 22

Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. and Tawfik, D. S. (December 2006).

Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein, *Nature*, **444**.7121, 929–932. 36

Bershtein, S. and Tawfik, D. S. (November 2008). Ohno's Model Revisited: Measuring the Frequency of Potentially Adaptive Mutations under Various Mutational Drifts, *Molecular Biology and Evolution*, **25**.11, 2311–2318. 36

Bieger, B. and Essen, L.-O. (March 2001). Crystal structure of the catalytic core component of the alkylhydroperoxide reductase AhpF from Escherichia coli, *Journal of Molecular Biology*, **307**.1, 1–8. 86, 88

Binkowski, T. A., Joachimiak, A. and Liang, J. (December 2005). Protein surface analysis for function annotation in high-throughput structural genomics pipeline, *Protein Science : a publication of the Protein Society*, **14**.12, 2972–2981. 46

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C. and Apweiler, R. (November 2009). QuickGO: a web-based tool for Gene Ontology searching., *Bioinformatics (Oxford, England)*, **25**.22, 3045–3046. 206

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. and Yuan, Y. (November 1998). Predicting function: from genes to genomes and back1, *Journal of Molecular Biology*, **283**.4, 707–725. 28

Brown, D. P., Krishnamurthy, N. and Sjölander, K. (August 2007). Automated Protein Subfamily Identification and Classification, *PLoS Computational Biology*, **3**.8, e160+. 41

Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S. and Kahn, D. (January 2005). The ProDom database of protein domain families: more emphasis on 3D, *Nucleic Acids Research*, **33**.suppl 1, D212–D215. 27

Buljan, M. and Bateman, A. (August 2009). The evolution of protein domain families., *Biochemical Society transactions*, **37**.Pt 4, 751–755. 105

Camps, M., Herman, A., Loh, E. and Loeb, L. A. (2007). Genetic constraints on protein evolution., *Critical Reviews in Biochemistry and Molecular Biology*, **42**.5, 313–326. 36

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pena, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J. and Knight, R. (May 2010). QIIME allows analysis of high-throughput community sequencing data, *Nature Methods*, **7**.5, 335–336. 171

Capra, J. A. and Singh, M. (July 2008). Characterization and prediction of residues determining protein functional specificity, *Bioinformatics (Oxford, England)*, **24**.13, 1473–1480. 42, 55, 185

Casari, G., Sander, C. and Valencia, A. (February 1995). A method to predict functional residues in proteins, *Nature Structural & Molecular Biology*, **2**.2, 171–178. 55

Caspi, R., Altman, T., Dale, J. M., Dreher, K., Fulcher, C. A., Gilham, F., Kaipa, P., Karthikeyan, A. S., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Paley, S., Popescu, L., Pujar, A., Shearer, A. G., Zhang, P. and Karp, P. D. (January 2010). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases., *Nucleic Acids Research*, **38**.Database issue, D473–D479. 29, 33

Celniker, G., Nimrod, G., Ashkenazy, H., Glaser, F., Martz, E., Mayrose, I., Pupko, T. and Ben-Tal, N. (April 2013). ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function, *Israel Journal of Chemistry*, **53**.3-4, 199–206. 54

Chakravorty, S., Helb, D., Burday, M., Connell, N. and Alland, D. (May 2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria., *Journal of Microbiological Methods*, **69**.2, 330–339. 165

Chen, C., Natale, D. A., Finn, R. D., Huang, H., Zhang, J., Wu, C. H. and Mazumder, R. (2011). Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation., *PLoS One*, **6**.4. 179

Chen, T., Yu, W.-H. H., Izard, J., Baranova, O. V., Lakshmanan, A. and Dewhirst, F. E. (January 2010). The Human Oral Microbiome Database: a web accessible resource for

investigating oral microbe taxonomic and genomic information., *Database : the journal of biological databases and curation*, **2010**.0, baq013+. 176

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Müller, W. E. G., Wetter, T. and Suhai, S. (June 2004). Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs, *Genome Research*, **14**.6, 1147–1159. 183

Chistoserdova, L. (October 2010). Recent progress and new challenges in metagenomics for biotechnology, *Biotechnology Letters*, **32**.10, 1351–1359. 166

Clarridge, J. E. (October 2004). Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases, *Clinical Microbiology Reviews*, **17**.4, 840–862. 165

Cooper, V. S. and Lenski, R. E. (October 2000). The population genetics of ecological specialization in evolving Escherichia coli populations., *Nature*, **407**.6805, 736–739. 36

Copley, S. D. (April 2003). Enzymes with extra talents: moonlighting functions and catalytic promiscuity., *Current Opinion in Chemical Biology*, **7**.2, 265–272. 108

Croft, D., Mundo, A. F. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L. and D'Eustachio, P. (January 2014). The Reactome pathway knowledgebase., *Nucleic Acids Research*, **42**.Database issue, D472–D477. 33

Cuff, A. L., Sillitoe, I., Lewis, T., Clegg, A. B., Rentzsch, R., Furnham, N., Pellegrini-Calace, M., Jones, D., Thornton, J. and Orengo, C. A. (January 2011). Extending CATH: increasing coverage of the protein structure universe and linking structure with function, *Nucleic Acids Research*, **39**.suppl 1, D420–D426. 41, 57, 59

Cuff, A. L., Sillitoe, I., Lewis, T., Redfern, O. C., Garratt, R., Thornton, J. and Orengo, C. A. (January 2009). The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies., *Nucleic Acids Research*, **37**.Database issue, D310–314. 122

Darveau, R. P. (June 2010). Periodontitis: a polymicrobial disruption of host homeostasis, *Nature Reviews Microbiology*, **8**.7, 481–490. 173

Dave, M., Higgins, P. D., Middha, S. and Rioux, K. P. (October 2012). The human gut microbiome: current knowledge, challenges, and future directions, *Translational Research*, **160**.4, 246–257. 175

de Spicer, P. O. and Maloy, S. (May 1993). PutA protein, a membrane-associated flavin dehydrogenase, acts as a redox-dependent transcriptional regulator., *Proceedings of the National Academy of Sciences of the United States of America*, **90**.9, 4295–4298. 109

Dekel, E. and Alon, U. (July 2005). Optimality and evolutionary tuning of the expression level of a protein, *Nature*, **436**.7050, 588–592. 36

Dellus-Gur, E., Toth-Petroczy, A., Elias, M. and Tawfik, D. S. (July 2013). What Makes a Protein Fold Amenable to Functional Innovation? Fold Polarity and Stability Trade-offs, *Journal of Molecular Biology*, **425**.14, 2609–2621. 100, 106, 107, 159, 163

den Besten, G., van Eunen, K., Groen, A. K., Venema, K., Reijngoud, D.-J. J. and Bakker, B. M. (September 2013). The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism., *Journal of Lipid Research*, **54**.9, 2325–2340. 215

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P. and Andersen, G. L. (July 2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB., *Applied and Environmental Microbiology*, **72**.7, 5069–5072. 171

Dessailly, B. H., Dawson, N. L., Mizuguchi, K. and Orengo, C. A. (March 2013). Functional site plasticity in domain superfamilies., *Biochimica et Biophysica Acta*. 59

Dessailly, B. H., Redfern, O. C., Cuff, A. and Orengo, C. A. (June 2009). Exploiting structural classifications for function prediction: towards a domain grammar for protein function., *Current Opinion in Structural Biology*, **19**.3, 349–356. 41, 45

Dessailly, B. H., Redfern, O. C., Cuff, A. L. and Orengo, C. A. (November 2010). Detailed analysis of function divergence in a large and diverse domain superfamily: toward a

refined protocol of function classification., *Structure (London, England : 1993)*, **18**.11, 1522–1535. 57, 62, 80, 96, 97, 104, 121

Devaraj, S., Hemarajata, P. and Versalovic, J. (April 2013). The human gut microbiome and body metabolism: implications for obesity and diabetes., *Clinical chemistry*, **59**.4, 617–628. 215

Dewhirst, F. E., Chen, T., Izard, J., Paster, B. J., Tanner, A. C., Yu, W.-H. H., Lakshmanan, A. and Wade, W. G. (October 2010). The human oral microbiome., *Journal of Bacteriology*, **192**.19, 5002–5017. 176, 218

Dinsdale, E. A., Edwards, R. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., Furlan, M., Desnues, C., Haynes, M., Li, L., McDaniel, L., Moran, M. A., Nelson, K. E., Nilsson, C., Olson, R., Paul, J., Brito, B. R., Ruan, Y., Swan, B. K., Stevens, R., Valentine, D. L., Thurber, R. V., Wegley, L., White, B. A. and Rohwer, F. (March 2008). Functional metagenomic profiling of nine biomes, *Nature*, **452**.7187, 629–632. 168

Drenth, J., Hol, W. G. J., Jansonius, J. N. and Koekoek, R. (January 1972). A Comparison of the Three-dimensional Structures of Subtilisin BPN and Subtilisin Novo, *Cold Spring Harbor Symposia on Quantitative Biology*, **36**, 107–116. 53, 109

Dubnau, D., Smith, I., Morell, P. and Marmur, J. (August 1965). Gene conservation in Bacillus species. I. Conserved genetic and nucleic acid base sequence homologies., *Proceedings of the National Academy of Sciences of the United States of America*, **54**.2, 491–498. 165

Duda, T. F. and Palumbi, S. R. (June 1999). Molecular genetics of ecological diversification: Duplication and rapid evolution of toxin genes of the venomous gastropod Conus, *Proceedings of the National Academy of Sciences of the United States of America*, **96**.12, 6820–6823. 36

Dutilh, B. E., Schmieder, R., Nulton, J., Felts, B., Salamon, P., Edwards, R. A. and Mokili, J. L. (December 2012). Reference-independent comparative metagenomics using cross-assembly: crAss., *Bioinformatics (Oxford, England)*, **28**.24, 3225–3231. 183

Dyda, F., Furey, W., Swaminathan, S., Sax, M., Farrenkopf, B. and Jordan, F. (June 1993). Catalytic centers in the thiamin diphosphate dependent enzyme pyruvate decarboxylase at 2.4-A resolution., *Biochemistry*, **32**.24, 6165–6170. 142

Easton, S., (2009). *Functional and Metagenomic Analysis of the Human Tongue Dorsum using Phage Display.* PhD thesis, University College London. 179

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. and Turner, S. (January 2009). Real-time DNA sequencing from single polymerase molecules., *Science (New York, N.Y.)*, **323**.5910, 133–138. 167

Ellrott, K., Jaroszewski, L., Li, W., Wooley, J. C. and Godzik, A. (June 2010). Expansion of the Protein Repertoire in Newly Explored Environments: Human Gut Microbiome Specific Protein Families, *PLoS Computational Biology*, **6**.6, e1000798+. 222

Enav, H., Mandel-Gutfreund, Y. and Béjà, O. (2014). Comparative metagenomic analyses reveal viral-induced shifts of host metabolism towards nucleotide biosynthesis., *Microbiome*, **2**.1. 187

Ewing, B. and Green, P. (March 1998). Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities, *Genome Research*, **8**.3, 186–194. 168

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J. and Punta, M. (January 2014). Pfam: the protein families database, *Nucleic Acids Research*, **42**.D1, D222–D230. 26, 27

Fischer, J. D., Holliday, G. L. and Thornton, J. M. (October 2010). The CoFactor database: organic cofactors in enzyme catalysis., *Bioinformatics (Oxford, England)*, **26**.19, 2496–2497. 100

Fleischmann, A., Darsow, M., Degtyarenko, K., Fleischmann, W., Boyce, S., Axelsen, K. B., Bairoch, A., Schomburg, D., Tipton, K. F. and Apweiler, R. (January 2004). IntEnz, the integrated relational enzyme database, *Nucleic Acids Research*, **32**.suppl 1, D434–D437. 32

Foerstner, K. U., Mering, C. and Bork, P. (March 2006). Comparative analysis of environmental sequences: potential and challenges, *Philosophical Transactions of the Royal Society B: Biological Sciences*, **361**.1467, 519–523. 164

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L. and Postlethwait, J. (April 1999). Preservation of duplicate genes by complementary, degenerative mutations., *Genetics*, **151**.4, 1531–1545. 34, 35

Foster, J. S., Palmer, R. J. and Kolenbrander, P. E. (April 2003). Human Oral Cavity as a Model for the Study of Genome-Genome Interactions, *The Biological Bulletin*, **204**.2, 200–204. 176

Furnham, N., Holliday, G. L., de Beer, T. A., Jacobsen, J. O., Pearson, W. R. and Thornton, J. M. (January 2014). The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes., *Nucleic Acids Research*, **42**.Database issue. 33, 34, 151

Furnham, N., Sillitoe, I., Holliday, G. L., Cuff, A. L., Rahman, S. A., Laskowski, R. A., Orengo, C. A. and Thornton, J. M. (January 2012). FunTree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies, *Nucleic Acids Research*, **40**.D1, D776–D782. 104

Gerlt, J. A. and Babbitt, P. C. (October 1998). Mechanistically diverse enzyme superfamilies: the importance of chemistry in the evolution of catalysis., *Current Opinion in Chemical Biology*, **2**.5, 607–612. 102, 103

Gherardini, P. F., Wass, M. N., Helmer-Citterich, M. and Sternberg, M. J. E. (September 2007). Convergent Evolution of Enzyme Active Sites Is not a Rare Phenomenon, *Journal of Molecular Biology*, **372**.3, 817–845. 109

Gibbons, R. J. and Houte, J. V. (1975). Bacterial Adherence in Oral Microbial Ecology, *Annual Review of Microbiology*, **29**.1, 19–42. 175, 176, 177

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M. and Nelson, K. E. (June 2006). Metagenomic Analysis of the Human Distal Gut Microbiome, *Science*, **312**.5778, 1355–1359. 173

Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E. and Ben-Tal, N. (January 2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information., *Bioinformatics (Oxford, England)*, **19**.1, 163–164. 54

Glasner, M., Gerlt, J. and Babbitt, P. (October 2006). Evolution of enzyme superfamilies, *Current Opinion in Chemical Biology*, **10**.5, 492–497. 37

Glenn, T. C. (September 2011). Field guide to next-generation DNA sequencers, *Molecular Ecology Resources*, **11**.5, 759–769. 167

Godzik, A. (June 2011). Metagenomics and the protein universe, *Current Opinion in Structural Biology*, **21**.3, 398–403. 222

Golovin, A., Dimitropoulos, D., Oldfield, T., Rachedi, A. and Henrick, K. (January 2005). MSDsite: a database search and retrieval system for the analysis and viewing of bound ligands and active sites., *Proteins*, **58**.1, 190–199. 46

Gough, J., Karplus, K., Hughey, R. and Chothia, C. (November 2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure, *Journal of Molecular Biology*, **313**.4, 903–919. 26

Greene, L. H., Lewis, T. E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A., Sillitoe, I., Yeats, C., Thornton, J. M. and Orengo, C. A. (January 2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution., *Nucleic Acids Research*, **35**.Database issue, D291–297. 24, 43

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. and Goodman, R. M. (October 1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products., *Chemistry & Biology*, **5**.10. 164

Hannenhalli, S. S. and Russell, R. B. (October 2000). Analysis and prediction of functional sub-types from protein sequence alignments, *Journal of Molecular Biology*, **303**.1, 61–76. 56

Harrison, A., Pearl, F., Sillitoe, I., Slidel, T., Mott, R., Thornton, J. and Orengo, C. (September 2003). Recognizing the fold of a protein structure., *Bioinformatics (Oxford, England)*, **19**.14, 1748–1759. 43

Hasson, M. S., Muscate, A., McLeish, M. J., Polovnikova, L. S., Gerlt, J. A., Kenyon, G. L., Petsko, G. A. and Ringe, D. (July 1998). The crystal structure of benzoylformate decarboxylase at 1.6 A resolution: diversity of catalytic residues in thiamin diphosphate-dependent enzymes., *Biochemistry*, **37**.28, 9918–9930. 142

Hawkins, T., Chitale, M., Luban, S. and Kihara, D. (February 2009). PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data, *Proteins*, **74**.3, 566–582. 39

Hildebrand, F., Meyer, A. and Eyre-Walker, A. (September 2010). Evidence of Selection upon Genomic GC-Content in Bacteria, *PLoS Genetics*, **6**.9, e1001107+. 193

Holliday, G. L., Andreini, C., Fischer, J. D., Rahman, S. A., Almonacid, D. E., Williams, S. T. and Pearson, W. R. (November 2011). MACiE: exploring the diversity of biochemical reactions, *Nucleic Acids Research*, **40**.Database issue, gkr799–D789. 32, 112, 113

Holm, L. and Sander, C. (September 1993). Protein structure comparison by alignment of distance matrices., *Journal of Molecular Biology*, **233**.1, 123–138. 44

Hooper, S. D., Dalevi, D., Pati, A., Mavromatis, K., Ivanova, N. N. and Kyrpides, N. C. (February 2010). Estimating DNA coverage and abundance in metagenomes using a gamma approximation, *Bioinformatics (Oxford, England)*, **26**.3, 295–301. 183

Hugenholtz, P., Goebel, B. M. and Pace, N. R. (September 1998). Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity, *Journal of Bacteriology*, **180**.18, 4765–4774. 164

Hugenholtz, P. and Tyson, G. W. (September 2008). Microbiology: Metagenomics, *Nature*, **455**.7212, 481–483. 168

Hughes, A. L., Green, J. A., Garbayo, J. M. and Roberts, R. M. (March 2000). Adaptive diversification within a large family of recently duplicated, placentally expressed genes., *Proceedings of the National Academy of Sciences of the United States of America*, **97**.7, 3319–3323. 36

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H. and Yeats, C. (January 2009). InterPro: the integrative protein signature database., *Nucleic Acids Research*, **37**.Database issue, D211–D215. 38

Hunter, S., Corbett, M., Denise, H., Fraser, M., Gonzalez-Beltran, A., Hunter, C., Jones, P., Leinonen, R., McAnulla, C., Maguire, E., Maslen, J., Mitchell, A., Nuka, G., Oisel, A., Pesseat, S., Radhakrishnan, R., Rocca-Serra, P., Scheremetjew, M., Sterk, P., Vaughan, D., Cochrane, G., Field, D. and Sansone, S.-A. (January 2014). EBI metagenomicsa new resource for the analysis and archiving of metagenomic data, *Nucleic Acids Research*, **42**.D1, D600–D606. 170, 171

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., de Castro, E., Coggill, P., Corbett, M., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Fraser, M., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., McMenamin, C., Mi, H., Mutowo-Muellenet, P., Mulder, N., Natale, D., Orengo, C., Pesseat, S., Punta, M., Quinn, A. F., Rivoire, C., Sangrador-Vegas, A., Selengut, J. D., Sigrist, C. J., Scheremetjew, M., Tate, J., Thimmajanarthanan, M., Thomas, P. D., Wu, C. H., Yeats, C. and Yong, S.-Y. Y. (January 2012). InterPro in 2011: new developments in the family and domain prediction database., *Nucleic Acids Research*, **40**.Database issue, D306–D312. 27, 40

Hunter, S. J., Easton, S., Booth, V., Henderson, B., Wade, W. G. and Ward, J. M. (2011). Selective removal of human DNA from metagenomic DNA samples extracted from dental plaque, *Journal of Basic Microbiology*, **51**.4, 442–446. 179

Huse, S., Huber, J., Morrison, H., Sogin, M. and Welch, D. (July 2007). Accuracy and quality of massively parallel DNA pyrosequencing, *Genome Biology*, **8**.7, R143+. 181

Jeffery, C. J. (January 1999). Moonlighting proteins., *Trends in Biochemical Sciences*, **24**.1, 8–11. 108

Jenkinson, H. F. (December 2011). Beyond the oral microbiome, *Environmental Microbiology*, **13**.12, 3077–3087. 176

Jensen, L. J. J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T. and Bork, P. (January 2008). eggNOG: automated construction and annotation of orthologous groups of genes., *Nucleic Acids Research*, **36**.Database issue, D250–254. 170

Jensen, R. A. (1976). Enzyme recruitment in evolution of new function., *Annual Review of Microbiology*, **30**, 409–425. 107

Jones, D. T., Taylor, W. R. and Thornton, J. M. (June 1992). The rapid generation of mutation data matrices from protein sequences., *Computer Applications in the Biosciences*, **8**.3, 275–282. 54

Jones, P., Binns, D., Chang, H.-Y. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y. Y., Lopez, R. and Hunter, S. (May 2014). InterProScan 5: genome-scale protein function classification., *Bioinformatics (Oxford, England)*, **30**.9, 1236–1240. 27

Kalinina, O. V., Novichkov, P. S., Mironov, A. A., Gelfand, M. S. and Rakhmaninova, A. B. (July 2004). SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins., *Nucleic Acids Research*, **32**.Web Server issue. 56

Kaminuma, E., Kosuge, T., Kodama, Y., Aono, H., Mashima, J., Gojobori, T., Sugawara, H., Ogasawara, O., Takagi, T., Okubo, K. and Nakamura, Y. (January 2011). DDBJ progress report, *Nucleic Acids Research*, **39**.suppl 1, D22–D27. 22

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (January 2014). Data, information, knowledge and principle: back to metabolism in KEGG., *Nucleic Acids Research*, **42**.Database issue, D199–D205. 29, 31, 33, 169, 216

Kaoutari, A. E., Armougom, F., Gordon, J. I., Raoult, D. and Henrissat, B. (June 2013). The abundance and variety of carbohydrate-active enzymes in the human gut microbiota, *Nature Reviews Microbiology*, **11**.7, 497–504. 192

Katoh, K., Kuma, K.-i., Toh, H. and Miyata, T. (January 2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment, *Nucleic Acids Research*, **33**.2, 511–518. 60

Katoh, K. and Toh, H. (July 2008). Recent developments in the MAFFT multiple sequence alignment program, *Briefings in Bioinformatics*, **9**.4, 286–298. 63

Khaladkar, M. and Hannenhalli, S. (July 2012). Functional divergence of gene duplicates - a domain-centric view, *BMC Evolutionary Biology*, **12**.1, 126+. 35

Khersonsky, O. and Tawfik, D. S. (2010). Enzyme promiscuity: a mechanistic and evolutionary perspective., *Annual Review of Biochemistry*, **79**, 471–505. 36, 37, 107, 108

Kim, M., Morrison, M. and Yu, Z. (January 2011). Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes., *Journal of Microbiological Methods*, **84**.1, 81–87. 165

Kinoshita, K. and Nakamura, H. (May 2004). eF-site and PDBjViewer: database and viewer for protein functional sites, *Bioinformatics*, **20**.8, 1329–1330. 45, 46

Kleywegt, G. J. (January 1999). Recognition of spatial motifs in protein structures., *Journal of Molecular Biology*, **285**.4, 1887–1897. 46

Knudsen, B. and Miyamoto, M. M. (December 2001). A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins., *Proceedings of the National Academy of Sciences of the United States of America*, **98**.25, 14512–14517. 56

Koehl, P., (February 2006). *Protein Structure Classification*, pages 1–55. John Wiley & Sons, Inc., Hoboken, NJ, USA. 44

Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J. and Lesk, A. M. (August 2006). MUSTANG: a multiple structural alignment algorithm., *Proteins*, **64**.3, 559–574. 106

Kosakovsky Pond, S., Wadhawan, S., Chiaromonte, F., Ananda, G., Chung, W.-Y., Taylor, J., Nekrutenko, A. and Team, T. G. (November 2009). Windshield splatter analysis with the Galaxy metagenomic pipeline, *Genome Research*, **19**.11, 2144–2153. 170

Koshland, D. E. (1958). Application of a Theory of Enzyme Specificity to Protein Synthesis, *Proceedings of the National Academy of Sciences of the United States of America*, **44**.2, 98–104. 99

Kotera, M., Okuno, Y., Hattori, M., Goto, S. and Kanehisa, M. (November 2004). Computational Assignment of the EC Numbers for Genomic-Scale Analysis of Enzymatic Reactions, *Journal of the American Chemical Society*, **126**.50, 16487–16498. 30

Kotrba, P., Inui, M. and Yukawa, H. (January 2001). Bacterial phosphotransferase system (PTS) in carbohydrate uptake and control of carbon metabolism, *Journal of Bioscience and Bioengineering*, **92**.6, 502–517. 216

Kraut, J. (1977). Serine Proteases: Structure and Mechanism of Catalysis, *Annual Review of Biochemistry*, **46**.1, 331–358. 53, 109

Krem, M. M., Rose, T. and Di Cera, E. (May 2000). Sequence determinants of function and evolution in serine proteases., *Trends in Cardiovascular Medicine*, **10**.4, 171–176. 100

Krissinel, E. and Henrick, K. (December 2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions., *Acta Crystallographica. Section D, Biological Crystallography*, **60**.Pt 12 Pt 1, 2256–2268. 43

Kristiansson, E., Hugenholtz, P. and Dalevi, D. (October 2009). ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes, *Bioinformatics*, **25**.20, 2737–2738. 164

Kruskal, W. H. (1957). Historical notes on the wilcoxon unpaired two-sample test, *Journal of the American Statistical Association*, **52**.279, pp. 356–360. 62

Kuriyan, J., Krishna, T. S. R., Wong, L., Guenther, B., Pahler, A., Williams, C. H. and Model, P. (July 1991). Convergent evolution of similar function in two structurally divergent enzymes, *Nature*, **352**.6331, 172–174. 110

Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., Takami, H., Morita, H., Sharma, V. K., Srivastava, T. P., Taylor, T. D., Noguchi, H., Mori, H., Ogura, Y., Ehrlich, D. S., Itoh, K., Takagi, T., Sakaki, Y., Hayashi, T. and Hattori, M. (August 2007). Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes., *DNA Research : an international journal for rapid publication of reports on genes and genomes*, **14**.4, 169–181. 215

La, D. and Livesay, D. R. (July 2005). MINER: software for phylogenetic motif identification, *Nucleic Acids Research*, **33**.suppl 2, W267–W270. 54

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J.,

Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J. and International Human Genome Sequencing Consortium (February 2001). Initial sequencing and analysis of the human genome., *Nature*, **409**.6822, 860–921. 166

Laskowski, R. A., Luscombe, N. M., Swindells, M. B. and Thornton, J. M. (December 1996). Protein clefts in molecular recognition and function., *Protein Science : a publication of the Protein Society*, **5**.12, 2438–2452. 45, 96, 97

Laskowski, R. A., Watson, J. D. and Thornton, J. M. (July 2005). ProFunc: a server for

predicting protein function from 3D structure., *Nucleic Acids Research*, **33**.Web Server issue, W89–W93. 47

Lee, D., Redfern, O. and Orengo, C. (December 2007). Predicting protein function from sequence and structure, *Nature Reviews Molecular Cell Biology*, **8**.12, 995–1005. 28, 37, 40, 43, 45, 46

Lee, D. A., Rentzsch, R. and Orengo, C. (January 2010). GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains., *Nucleic Acids Research*, **38**.3, 720–737. 41, 59, 60

Lees, J., Yeats, C., Perkins, J., Sillitoe, I., Rentzsch, R., Dessailly, B. H. and Orengo, C. (January 2012). Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis., *Nucleic Acids Research*, **40**.Database issue, D465–D471. 59

Lees, J. G., Lee, D., Studer, R. A., Dawson, N. L., Sillitoe, I., Das, S., Yeats, C., Dessailly, B. H., Rentzsch, R. and Orengo, C. A. (January 2014). Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis, *Nucleic Acids Research*, **42**.D1, D240–D245. 23, 26, 93

Leininger, S., Urich, T., Schloter, M., Schwark, L., Qi, J., Nicol, G. W., Prosser, J. I., Schuster, S. C. and Schleper, C. (August 2006). Archaea predominate among ammonia-oxidizing prokaryotes in soils, *Nature*, **442**.7104, 806–809. 168

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Hoopen, P. T., Vaughan, R., Zalunin, V. and Cochrane, G. (January 2011). The European Nucleotide Archive, *Nucleic Acids Research*, **39**.suppl 1, D28–D31. 22

Letunic, I., Doerks, T. and Bork, P. (January 2009). SMART 6: recent updates and new developments, *Nucleic Acids Research*, **37**.suppl 1, D229–D232. 27

Li, H. and Durbin, R. (March 2010). Fast and accurate long-read alignment with Burrows-Wheeler transform., *Bioinformatics (Oxford, England)*, **26**.5, 589–595. 182

Li, W. and Godzik, A. (July 2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, **22**.13, 1658–1659. 60

Lichtarge, O., Bourne, H. R. and Cohen, F. E. (March 1996). An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families, *Journal of Molecular Biology*, **257**.2, 342–358. 54, 55

Lima, T., Auchincloss, A. H., Coudert, E., Keller, G., Michoud, K., Rivoire, C., Bulliard, V., de Castro, E., Lachaize, C., Baratin, D., Phan, I., Bougueleret, L. and Bairoch, A. (January 2009). HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot, *Nucleic Acids Research*, **37**.suppl 1, D471–D478. 27

Lindqvist, Y. (September 1989). Refined structure of spinach glycolate oxidase at 2 A resolution., *Journal of Molecular Biology*, **209**.1, 151–166. 156

Litvak, Y. and Selinger, Z. (December 2003). Bacterial mimics of eukaryotic GTPase-activating proteins (GAPs)., *Trends in Biochemical Sciences*, **28**.12, 628–631. 82

Livingstone, C. D. and Barton, G. J. (December 1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation., *Computer Applications in the Biosciences*, **9**.6, 745–756. 53, 54

Lynch, M., Walsh, B. and Others (1998). Genetics and analysis of quantitative traits. 34

Lysholm, F., Wetterbom, A., Lindau, C., Darban, H., Bjerkner, A., Fahlander, K., Lindberg, A. M., Persson, B., Allander, T. and Andersson, B. (February 2012). Characterization of the Viral Microbiome in Patients with Severe Lower Respiratory Tract Infections, Using Metagenomic Sequencing, *PLoS ONE*, **7**.2, e30875+. 183

Macheroux, P., Kieweg, V., Massey, V., Söderlind, E., Stenberg, K. and Lindqvist, Y. (May 1993). Role of tyrosine 129 in the active site of spinach glycolate oxidase., *European Journal of Biochemistry / FEBS*, **213**.3, 1047–1054. 155

Maeda-Yorita, K., Aki, K., Sagai, H., Misaki, H. and Massey, V. (1995). L-lactate oxidase and L-lactate monooxygenase: mechanistic variations on a common structural theme., *Biochimie*, **77**.7-8, 631–642. 155

Maiorov, V. N. and Crippen, G. M. (January 1994). Significance of Root-Mean-Square Deviation in Comparing Three-dimensional Structures of Globular Proteins, *Journal of Molecular Biology*, **235**.2, 625–634. 44

Makarova, K. S. and Grishin, N. V. (September 1999). The Zn-peptidase superfamily: functional convergence after evolutionary divergence., *Journal of Molecular Biology*, **292**.1, 11–17. 111

Manning, J., Jefferson, E. and Barton, G. (2008). The contrasting properties of conservation and correlated phylogeny in protein functional residue prediction, *BMC Bioinformatics*, **9**.1, 51+. 54

Mardis, E. R. (2013). Next-generation sequencing platforms., *Annual Review of Analytical Chemistry*, **6**, 287–303. 167

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H. H., Ho, C. H. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J.-B. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. and Rothberg, J. M. (September 2005). Genome sequencing in microfabricated high-density picolitre reactors., *Nature*, **437**.7057, 376–380. 167

Markowitz, V. M., Chen, I.-M. A., Chu, K., Szeto, E., Palaniappan, K., Grechkin, Y., Ratner, A., Jacob, B., Pati, A., Huntemann, M., Liolios, K., Pagani, I., Anderson, I., Mavromatis, K., Ivanova, N. N. and Kyrpides, N. C. (January 2012). IMG/M: the integrated metagenome data management and comparative analysis system, *Nucleic Acids Research*, **40**.D1, D123–D129. 170, 171

Marsh, P. D. and Devine, D. A. (March 2011). How is the development of dental biofilms

influenced by the host?, *Journal of Clinical Periodontology*, **38 Suppl 11**, 28–35. 176, 177

Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., Mitchell, J. B., Taroni, C. and Thornton, J. M. (July 1998). Protein folds and functions., *Structure (London, England : 1993)*, **6**.7, 875–884. 44

Martin, D. M., Berriman, M. and Barton, G. J. (November 2004). GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes., *BMC Bioinformatics*, **5**.1, 178+. 38

Maxam, A. M. and Gilbert, W. (February 1977). A new method for sequencing DNA., *Proceedings of the National Academy of Sciences of the United States of America*, **74**.2, 560–564. 166

McDonald, A. G., Boyce, S. and Tipton, K. F. (January 2009). ExplorEnz: the primary source of the IUBMB enzyme list., *Nucleic Acids Research*, **37**.Database issue, gkn582+. 31

McLachlan, A. D. (March 1972). Repeating sequences and gene duplication in proteins, *Journal of Molecular Biology*, **64**.2, 417–437. 115, 117

McLoughlin, S. Y. and Ollis, D. L. (June 2004). The Role of Inhibition in Enzyme Evolution, *Chemistry & Biology*, **11**.6, 735–737. 36

Metzker, M. L. (January 2010). Sequencing technologies - the next generation, *Nature Reviews Genetics*, **11**.1, 31–46. 168

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J. and Edwards, R. A. (December 2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes, *BMC Bioinformatics*, **9**.1, 386–8. 170, 171, 172, 187

Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S. and Thomas, P. D. (January 2010). PANTHER version 7: improved phylogenetic trees, orthologs and collaboration

with the Gene Ontology Consortium, *Nucleic Acids Research*, **38**.suppl 1, D204–D210. 27

Mihalek, I., Reš, I. and Lichtarge, O. (March 2004). A Family of EvolutionEntropy Hybrid Methods for Ranking Protein Residues by Importance, *Journal of Molecular Biology*, **336**.5, 1265–1282. 53

Miller, J. R., Koren, S. and Sutton, G. (June 2010). Assembly algorithms for next-generation sequencing data, *Genomics*, **95**.6, 315–327. 172

Mirny, L. A. and Gelfand, M. S. (August 2002). Using Orthologous and Paralogous Proteins to Identify Specificity-determining Residues in Bacterial Transcription Factors, *Journal of Molecular Biology*, **321**.1, 7–20. 56

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. and Punta, M. (July 2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions, *Nucleic Acids Research*, **41**.12, e121. 185

Mitra, S., Klar, B. and Huson, D. H. (August 2009). Visual and statistical comparison of metagenomes, *Bioinformatics (Oxford, England)*, **25**.15, 1849–1855. 221

Muhammadzadeh, A., Trost, B., Pittet, V., Johnson, S. and Kusalik, A., (2013). Comparing de novo assemblers for metagenomic data. F1000 posters. 183

Muro-Pastor, A. M., Ostrovsky, P. and Maloy, S. (April 1997). Regulation of gene expression by repressor localization: biochemical evidence that membrane and DNA binding by the PutA protein are mutually exclusive., *Journal of Bacteriology*, **179**.8, 2788–2791. 109

Mutz, K.-O., Heilkenbrinker, A., Lönne, M., Walter, J.-G. and Stahl, F. (February 2013). Transcriptome analysis using next-generation sequencing, *Current Opinion in Biotechnology*, **24**.1, 22–30. 166

Nagarajan, N. and Pop, M. (March 2013). Sequence assembly demystified, *Nature Reviews Genetics*, **14**.3, 157–167. 166, 167, 168

Neidhart, D. J., Kenyon, G. L., Gerlt, J. A. and Petsko, G. A. (October 1990). Mandelate racemase and muconate lactonizing enzyme are mechanistically distinct and structurally homologous, *Nature*, **347**.6294, 692–694. 101

NIH HMP Working Group, Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., Bonazzi, V., McEwen, J. E., Wetterstrand, K. A., Deal, C., Baker, C. C., Di Francesco, V., Howcroft, T. K., Karp, R. W., Lunsford, R. D., Wellington, C. R., Belachew, T., Wright, M., Giblin, C., David, H., Mills, M., Salomon, R., Mullins, C., Akolkar, B., Begg, L., Davis, C., Grandison, L., Humble, M., Khalsa, J., Little, A. R., Peavy, H., Pontzer, C., Portnoy, M., Sayre, M. H., Starke-Reed, P., Zakhari, S., Read, J., Watson, B. and Guyer, M. (December 2009). The NIH Human Microbiome Project., *Genome Research*, **19**.12, 2317–2323. 174

Nikolskayaw, Q. N., Arighi, C. N., Huang, H., Barker, W. C. and Wu, C. H. (2006). PIRSF Family Classification System for Protein Functional and Evolutionary Analysis, *Evolutionary Bioinformatics*, **2**. 27

O'Boyle, N. M., Holliday, G. L., Almonacid, D. E. and Mitchell, J. B. O. (May 2007). Using Reaction Mechanism to Measure Enzyme Similarity, *Journal of Molecular Biology*, **368**.5, 1484–1499. 32, 110

Ohno, S., (1970). *Evolution by gene duplication*. Springer-Verlag, New York, Heidelberg, Berlin. 34

Ojha, S., Meng, E. C. and Babbitt, P. C. (July 2007). Evolution of Function in the "Two Dinucleotide Binding Domains" Flavoproteins, *PLoS Comput Biol*, **3**.7, e121+. 85

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. and Thornton, J. M. (August 1997). CATH a hierarchic classification of protein domain structures, *Structure*, **5**.8, 1093–1109. 24, 25

Orengo, C. A. and Taylor, W. R., (1996). *[36] SSAP: Sequential structure alignment program for protein structure comparison*, volume 266 of *Methods in Enzymology*, pages 617–635. Elsevier. 43, 122

Pagani, I., Liolios, K., Jansson, J., Chen, I.-M. M., Smirnova, T., Nosrat, B., Markowitz, V. M. and Kyrpides, N. C. (January 2012). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata., *Nucleic Acids Research*, **40**.Database issue, D571–D579. 21

Pal, D. and Eisenberg, D. (January 2005). Inference of protein function from protein structure., *Structure (London, England : 1993)*, **13**.1, 121–130. 47, 48

Pegg, S. C.-H. C., Brown, S. D., Ojha, S., Seffernick, J., Meng, E. C., Morris, J. H., Chang, P. J., Huang, C. C., Ferrin, T. E. and Babbitt, P. C. (February 2006). Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database., *Biochemistry*, **45**.8, 2545–2555. 48, 60

Petrosino, J. F., Highlander, S., Luna, R. A., Gibbs, R. A. and Versalovic, J. (May 2009). Metagenomic Pyrosequencing and Microbial Identification, *Clinical Chemistry*, **55**.5, 856–866. 166

Petsko, G. A., Kenyon, G. L., Gerlt, J. A., Ringe, D. and Kozarich, J. W. (October 1993). On the origin of enzymatic species, *Trends in Biochemical Sciences*, **18**.10, 372–376. 102

Poinar, H. N., Schwarz, C., Qi, J., Shapiro, B., Macphee, R. D., Buigues, B., Tikhonov, A., Huson, D. H., Tomsho, L. P., Auch, A., Rampp, M., Miller, W. and Schuster, S. C. (January 2006). Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA., *Science (New York, N.Y.)*, **311**.5759, 392–394. 168

Polovnikova, E. S., McLeish, M. J., Sergienko, E. A., Burgner, J. T., Anderson, N. L., Bera, A. K., Jordan, F., Kenyon, G. L. and Hasson, M. S. (January 2003). Structural and Kinetic Analysis of Catalysis by a Thiamin Diphosphate-Dependent Enzyme, Benzoylformate Decarboxylase†, *Biochemistry*, **42**.7, 1820–1830. 142

Porter, C. T., Bartlett, G. J. and Thornton, J. M. (January 2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data., *Nucleic Acids Research*, **32**.Database issue, D129–D133. 33, 46, 61, 64, 101, 114

Porter, J. R. (June 1976). Antony van Leeuwenhoek: tercentenary of his discovery of bacteria., *Bacteriological Reviews*, **40**.2, 260–269. 175

Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., Gabaldón, T., Rattei, T., Creevey, C., Kuhn, M., Jensen, L. J., von Mering, C. and Bork, P. (January 2014). eggNOG v4.0: nested orthology inference across 3686 organisms, *Nucleic Acids Research*, **42**.D1, D231–D239. 170

Powers, R., Copeland, J. C., Germer, K., Mercier, K. A., Ramanathan, V. and Revesz, P. (October 2006). Comparison of protein active site structures for functional annotation of proteins and drug design, *Proteins*, **65**.1, 124–135. 47

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J. and Glöckner, F. O. O. (December 2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB., *Nucleic Acids Research*, **35**.21, 7188–7196. 171

Pruitt, K. D., Tatusova, T. and Maglott, D. R. (January 2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins., *Nucleic Acids Research*, **33**.Database issue, D501–D504. 22

Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. and Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues., *Bioinformatics (Oxford, England)*, **18 Suppl 1**. 53

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M. M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., MetaHIT Consortium, Bork, P., Ehrlich, S. D. and Wang, J. (March 2010). A human gut microbial gene catalogue established by metagenomic sequencing., *Nature*, **464**.7285, 59–65. 168, 174

R Core Team, (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. 62

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., Pandey, G., Yunes, J. M., Talwalkar, A. S., Repo, S., Souza, M. L., Piovesan, D., Casadio, R., Wang, Z., Cheng, J., Fang, H., Gough, J., Koskinen, P., Toronen, P., Nokso-Koivisto, J., Holm, L., Cozzetto, D., Buchan, D. W. A., Bryson, K., Jones, D. T., Limaye, B., Inamdar, H., Datta, A., Manjari, S. K., Joshi, R., Chitale, M., Kihara, D., Lisewski, A. M., Erdin, S., Venner, E., Lichtarge, O., Rentzsch, R., Yang, H., Romero, A. E., Bhat, P., Paccanaro, A., Hamp, T., Kaszner, R., Seemayer, S., Vicedo, E., Schaefer, C., Achten, D., Auer, F., Boehm, A., Braun, T., Hecht, M., Heron, M., Honigschmid, P., Hopf, T. A., Kaufmann, S., Kiening, M., Krompass, D., Landerer, C., Mahlich, Y., Roos, M., Bjorne, J., Salakoski, T., Wong, A., Shatkay, H., Gatzmann, F., Sommer, I., Wass, M. N., Sternberg, M. J. E., Skunca, N., Supek, F., Bosnjak, M., Panov, P., Dzeroski, S., Smuc, T., Kourmpetis, Y. A. I., van Dijk, A. D. J., Braak, C. J., Zhou, Y., Gong, Q., Dong, X., Tian, W., Falda, M., Fontana, P., Lavezzo, E., Di Camillo, B., Toppo, S., Lan, L., Djuric, N., Guo, Y., Vucetic, S., Bairoch, A., Linial, M., Babbitt, P. C., Brenner, S. E., Orengo, C., Rost, B., Mooney, S. D. and Friedberg, I. (January 2013). A large-scale evaluation of computational protein function prediction, *Nature Methods*, **10**.3, 221–227. 37, 49, 50, 51, 61

Raes, J., Foerstner, K. U. U. and Bork, P. (October 2007). Get the most out of your metagenome: computational analysis of environmental sequence data., *Current Opinion in Microbiology*, **10**.5, 490–498. 173

Rahman, S. A., Cuesta, S. M., Furnham, N., Holliday, G. L. and Thornton, J. M. (January 2014). EC-BLAST: a tool to automatically search and compare enzyme reactions, *Nature Methods*, **11**.2, 171–174. 111, 120

Rao, S. T. and Rossmann, M. G. (May 1973). Comparison of super-secondary structures in proteins., *Journal of Molecular Biology*, **76**.2, 241–256. 44

Rausell, A., Juan, D., Pazos, F. and Valencia, A. (January 2010). Protein interactions

and ligand binding: From protein subfamilies to functional specificity, *Proceedings of the National Academy of Sciences of the United States of America*, **107**.5, 1995–2000. 62

Redfern, O., Dessailly, B. and Orengo, C. (June 2008). Exploring the structure and function paradigm, *Current Opinion in Structural Biology*, **18**.3, 394–402. 44, 76, 121

Reeves, G., Dallman, T., Redfern, O., Akpor, A. and Orengo, C. (July 2006). Structural Diversity of Domain Superfamilies in the CATH Database, *Journal of Molecular Biology*, **360**.3, 725–741. 105

Reid, A., Ranea, J. and Orengo, C. (February 2010). Comparative evolutionary analysis of protein complexes in E. coli and yeast, *BMC Genomics*, **11**.1, 79+. 80, 97

Rentzsch, R. and Orengo, C. (2012). Protein function prediction using domain families, *BMC Bioinformatics*. 60, 61

Rho, M., Tang, H. and Ye, Y. (November 2010). FragGeneScan: predicting genes in short and error-prone reads., *Nucleic Acids Research*, **38**.20, e191. 171

Rice, P., Longden, I. and Bleasby, A. (June 2000). EMBOSS: the European Molecular Biology Open Software Suite., *Trends in genetics : TIG*, **16**.6, 276–277. 183, 185

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. and Nyrén, P. (November 1996). Real-Time DNA Sequencing Using Detection of Pyrophosphate Release, *Analytical Biochemistry*, **242**.1, 84–89. 166

Rost, B. (2002). Enzyme function less conserved than anticipated, *Journal of Molecular Biology*, **318**, 595–608. 38

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Güldener, U., Mannhaupt, G., Münsterkötter, M. and Mewes, H. W. (January 2004). The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes, *Nucleic Acids Research*, **32**.18, 5539–5545. 28, 30

Russell, R. B. and Barton, G. J. (October 1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels., *Proteins*, **14**.2, 309–323. 43

Sadowski, M. I. and Jones, D. T. (June 2009). The sequencestructure relationship and protein function prediction, *Current Opinion in Structural Biology*, **19**.3, 357–362. 40

Sadreyev, R. and Grishin, N. (February 2003). COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance., *Journal of Molecular Biology*, **326**.1, 317–336. 60

Sandhya, S., Rani, S. S. S., Pankaj, B., Govind, M. K. K., Offmann, B., Srinivasan, N. and Sowdhamini, R. (2009). Length variations amongst protein domain superfamilies and consequences on structure and function., *PLoS One*, **4**.3. 105

Sanger, F., Nicklen, S. and Coulson, A. R. (December 1977). DNA sequencing with chain-terminating inhibitors, *Proceedings of the National Academy of Sciences of the United States of America*, **74**.12, 5463–5467. 166

Schmidt, T. M., DeLong, E. F. and Pace, N. R. (July 1991). Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing., *Journal of Bacteriology*, **173**.14, 4371–4378. 165

Schmieder, R. and Edwards, R. (March 2011a). Fast identification and removal of sequence contamination from genomic and metagenomic datasets., *PLoS One*, **6**.3, e17288+. 182

Schmieder, R. and Edwards, R. (March 2011b). Quality control and preprocessing of metagenomic datasets, *Bioinformatics (Oxford, England)*, **27**.6, 863–864. 181, 183

Schomburg, I., Chang, A., Placzek, S., Söhngen, C., Rother, M., Lang, M., Munaretto, C., Ulas, S., Stelzer, M., Grote, A., Scheer, M. and Schomburg, D. (January 2013). BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA., *Nucleic Acids Research*, **41**.Database issue. 31

Selengut, J. D., Haft, D. H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W. C., Richter, A. R. and White, O. (January 2007). TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes, *Nucleic Acids Research*, **35**.suppl 1, D260–D264. 27

Shannon, C. E. and Weaver, W., (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana. 53

Shestakov, S. (March 2011). Metagenomics of the human microbiome, *Biology Bulletin Reviews*, **1**.2, 83–93. 176

Shindyalov, I. N. and Bourne, P. E. (September 1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path., *Protein Engineering*, **11**.9, 739–747. 43

Shoemaker, B. A., Zhang, D., Tyagi, M., Thangudu, R. R., Fong, J. H., Marchler-Bauer, A., Bryant, S. H., Madej, T. and Panchenko, A. R. (January 2012). IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins., *Nucleic Acids Research*, **40**.Database issue, D834–D840. 62, 64

Sierk, M. L. (February 2004). Sensitivity and selectivity in protein structure comparison, *Protein Science : a publication of the Protein Society*, **13**.3, 773–785. 34

Sigrist, C. J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A. and Bucher, P. (September 2002). PROSITE: a documented database using patterns and profiles as motif descriptors., *Briefings in Bioinformatics*, **3**.3, 265–274. 40

Sigrist, C. J. A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A. and Hulo, N. (January 2010). PROSITE, a protein domain database for functional characterization and annotation, *Nucleic Acids Research*, **38**.suppl 1, D161–D166. 27

Sillitoe, I. and Orengo, C., (November 2002). Protein Structure Comparison. In Orengo, C., Jones, D. and Thornton, J. (eds.), *Bioinformatics: Genes, Proteins and Computers*, chapter 6, pages 79–99. Taylor & Francis. 44

Sillitoe, I., Cuff, A. L., Dessailly, B. H., Dawson, N. L., Furnham, N., Lee, D., Lees, J. G., Lewis, T. E., Studer, R. A., Rentzsch, R., Yeats, C., Thornton, J. M. and Orengo, C. A. (January 2013). New functional families (FunFams) in CATH to improve the mapping

of conserved functional sites to 3D structures., *Nucleic Acids Research*, **41**.Database issue, D490–D498. 23, 24, 25, 41, 69, 93

Sjölander, K. (1998). Phylogenetic inference in protein superfamilies: analysis of SH2 domains., *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, **6**, 165–174. 41, 56

Söding, J. (April 2005). Protein homology detection by HMM-HMM comparison., *Bioinformatics (Oxford, England)*, **21**.7, 951–960. 40

Sonnhammer, E. L., Eddy, S. R. and Durbin, R. (July 1997). Pfam: a comprehensive database of protein domain families based on seed alignments., *Proteins*, **28**.3, 405–420. 41

Sonnhammer, E. L. L. and Koonin, E. V. (December 2002). Orthology, paralogy and proposed classification for paralog subtypes, *Trends in Genetics*, **18**.12, 619–620. 24

Spraggon, G., Kim, C., Nguyen-Huu, X., Yee, M. C., Yanofsky, C. and Mills, S. E. (May 2001). The structures of anthranilate synthase of Serratia marcescens crystallized in the presence of (i) its substrates, chorismate and glutamine, and a product, glutamate, and (ii) its end-product inhibitor, L-tryptophan., *Proceedings of the National Academy of Sciences of the United States of America*, **98**.11, 6021–6026. 153

Spriggs, R. V., Artymiuk, P. J. and Willett, P. (2003). Searching for patterns of amino acids in 3D protein structures., *Journal of Chemical Information and Computer Sciences*, **43**.2, 412–421. 46

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D. and Birney, E. (October 2002). The Bioperl toolkit: Perl modules for the life sciences., *Genome Research*, **12**.10, 1611–1618. 120

Stoebel, D. M., Dean, A. M. and Dykhuizen, D. E. (March 2008). The Cost of Expression

of Escherichia coli lac Operon Proteins Is in the Process, Not in the Products, *Genetics*, **178**.3, 1653–1660. 36

Subbiah, S., Laurents, D. V. and Levitt, M. (March 1993). Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core, *Current Biology*, **3**.3, 141–148. 43

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J. and Natale, D. A. (September 2003). The COG database: an updated version includes eukaryotes., *BMC Bioinformatics*, **4**.1, 41+. 28, 169

Taylor, W. R. (March 1986). The classification of amino acid conservation, *Journal of Theoretical Biology*, **119**.2, 205–218. 53

Taylor, W. R. and Orengo, C. A. (July 1989). Protein structure alignment, *Journal of Molecular Biology*, **208**.1, 1–22. 43, 63, 64, 65, 66, 115

The UniProt Consortium (January 2014). Activities at the Universal Protein Resource (UniProt), *Nucleic Acids Research*, **42**.D1, D191–D198. 22, 30

Thoden, J. B., Holden, H. M., Wesenberg, G., Raushel, F. M. and Rayment, I. (May 1997). Structure of Carbamoyl Phosphate Synthetase: A Journey of 96 Å from Substrate to Product†,‡, *Biochemistry*, **36**.21, 6305–6316. 152

Thoden, J. B., Wesenberg, G., Raushel, F. M. and Holden, H. M. (February 1999). Carbamoyl Phosphate Synthetase: Closure of the B-Domain as a Result of Nucleotide Binding†,‡, *Biochemistry*, **38**.8, 2347–2357. 153

Thompson, A. D., Dugan, A., Gestwicki, J. E. and Mapp, A. K. (June 2012). Fine-Tuning Multiprotein Complexes Using Small Molecules, *American Chemical Society Chemical Biology*, **7**.8, 1311–1320. 97, 98

Todd, A. E., Orengo, C. A. and Thornton, J. M. (2002). Sequence and structural differences between enzyme and nonenzyme homologs., *Structure*, **10**, 1435–1451. 110

Todd, A. E., Orengo, C. A. and Thornton, J. M. (April 2001). Evolution of function in protein superfamilies, from a structural perspective, *Journal of Molecular Biology*, **307**.4, 1113–1143. 80, 103, 104, 144, 162

Todd, A. E., Orengo, C. A. and Thornton, J. M. (1999). Evolution of protein function, from a structural perspective, *Current Opinion in Chemical Biology*, **3**.5, 548 – 556. 105

Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L. and Tawfik, D. S. (June 2007). The stability effects of protein mutations appear to be universally distributed., *Journal of Molecular Biology*, **369**.5, 1318–1332. 36

Tokuriki, N. and Tawfik, D. S. (October 2009). Stability effects of mutations and protein evolvability, *Current Opinion in Structural Biology*, **19**.5, 596–604. 36

Torrance, J. W., Bartlett, G. J., Porter, C. T. and Thornton, J. M. (April 2005). Using a Library of Structural Templates to Recognise Catalytic Sites and Explore their Evolution in Homologous Families, *Journal of Molecular Biology*, **347**.3, 565–581. 46, 47

Trimble, W., Keegan, K., D'Souza, M., Wilke, A., Wilkening, J., Gilbert, J. and Meyer, F. (July 2012). Short-read reading-frame predictors are not created equal: sequence error causes loss of signal, *BMC Bioinformatics*, **13**.1, 183+. 184

Tse, H., Tsang, A. K. L., Tsoi, H.-W., Leung, A. S. P., Ho, C.-C., Lau, S. K. P., Woo, P. C. Y. and Yuen, K.-Y. (August 2012). Identification of a Novel Bat Papillomavirus by Metagenomics, *PLoS ONE*, **7**.8, e43986+. 183

Turnbaugh, P. J. and Gordon, J. I. (September 2008). An Invitation to the Marriage of Metagenomics and Metabolomics, *Cell*, **134**.5, 708–713. 173, 174

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R. and Gordon, J. I. (October 2007). The Human Microbiome Project, *Nature*, **449**.7164, 804–810. 174

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S. and Banfield, J. F. (March 2004). Com-

munity structure and metabolism through reconstruction of microbial genomes from the environment., *Nature*, **428**.6978, 37–43. 164

Valdar, W. S. J. (August 2002). Scoring residue conservation, *Proteins*, **48**.2, 227–241. 54, 62, 63

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H. and Smith, H. O. (April 2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea, *Science*, **304**.5667, 66–74. 164

Wade, W. G. (March 2013). The oral microbiome in health and disease, *Pharmacological Research*, **69**.1, 137–143. 175, 176, 177, 217, 219

Wagner, A. (June 2005). Energy constraints on the evolution of gene expression., *Molecular Biology and Evolution*, **22**.6, 1365–1374. 36

Wang, Q., Garrity, G. M., Tiedje, J. M. and Cole, J. R. (August 2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy, *Applied and Environmental Microbiology*, **73**.16, 5261–5267. 171

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. and Barton, G. J. (May 2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench., *Bioinformatics (Oxford, England)*, **25**.9, 1189–1191. 137, 143

Watson, J. D., Laskowski, R. A. and Thornton, J. M. (June 2005). Predicting protein function from sequence and structural data, *Current Opinion in Structural Biology*, **15**.3, 275–284. 46

Wegley, L., Edwards, R., Rodriguez-Brito, B., Liu, H. and Rohwer, F. (November 2007). Metagenomic analysis of the microbial community associated with the coral Porites astreoides., *Environmental Microbiology*, **9**.11, 2707–2719. 168

Wickström, C., Herzberg, M. C., Beighton, D. and Svensäter, G. (September 2009). Proteolytic degradation of human salivary MUC5B by dental biofilms., *Microbiology (Reading, England)*, **155**.Pt 9, 2866–2872. 177

Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C. and Gough, J. (January 2009). SUPERFAMILYsophisticated comparative genomics, data mining, visualization and phylogeny, *Nucleic Acids Research*, **37**.suppl 1, D380–D386. 23, 26

Wilson, M., (2004). *Microbial Inhabitants of Humans*. Cambridge University Press, Cambridge. 175

Wittig, U., Kania, R., Golebiewski, M., Rey, M., Shi, L., Jong, L., Algaa, E., Weidemann, A., Sauer-Danzwith, H., Mir, S., Krebs, O., Bittkowski, M., Wetsch, E., Rojas, I. and Müller, W. (January 2012). SABIO-RK–database for biochemical reaction kinetics., *Nucleic Acids Research*, **40**.Database issue. 32

Woese, C. R. (June 1987). Bacterial evolution., *Microbiological Reviews*, **51**.2, 221–271. 165

Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Mazumder, R., O'Donovan, C., Redaschi, N. and Suzek, B. (January 2006). The Universal Protein Resource (UniProt): an expanding universe of protein information, *Nucleic Acids Research*, **34**.suppl 1, D187–D191. 28

Wu, D., Doroud, L. and Eisen, J. A., (August 2013). TreeOTU: Operational Taxonomic Unit Classification Based on Phylogenetic Trees. 169

Wu, T. T. and Kabat, E. A. (August 1970). An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity., *The Journal of Experimental Medicine*, **132**.2, 211–250. 53

Xia, Z. X., Shamala, N., Bethge, P. H., Lim, L. W., Bellamy, H. D., Xuong, N. H., Lederer, F. and Mathews, F. S. (May 1987). Three-dimensional structure of flavocytochrome b2 from baker's yeast at 3.0-A resolution., *Proceedings of the National Academy of Sciences of the United States of America*, **84**.9, 2629–2633. 156

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C.,

Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, C. A., Lauber, C., Clemente, J. C., Knights, D., Knight, R. and Gordon, J. I. (May 2012). Human gut microbiome viewed across age and geography., *Nature*, **486**.7402, 222–227. 179, 219

Yeats, C., Redfern, O. C. and Orengo, C. (March 2010). A fast and automated solution for accurately resolving protein domain architectures, *Bioinformatics*, **26**.6, 745–751. 185

Yue, P., Li, Z. and Moult, J. (October 2005). Loss of Protein Structure Stability as a Major Causative Factor in Monogenic Disease, *Journal of Molecular Biology*, **353**.2, 459–473. 36

Zhang, J., Rosenberg, H. F. and Nei, M. (March 1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes, *Proceedings of the National Academy of Sciences of the United States of America*, **95**.7, 3708–3713. 36

Zhang, M., Van Etten, R. L. and Stauffacher, C. V. (September 1994). Crystal structure of bovine heart phosphotyrosyl phosphatase at 2.2-A resolution., *Biochemistry*, **33**.37, 11097–11105. 109

Zhu, W., Lomsadze, A. and Borodovsky, M. (July 2010). Ab initio gene identification in metagenomic sequences, *Nucleic Acids Research*, **38**.12, e132. 184

Zvelebil, M. J., Barton, G. J., Taylor, W. R. and Sternberg, M. J. E. (June 1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences, *Journal of Molecular Biology*, **195**.4, 957–961. 53